Calibrating a stochastic SIR-model to simulated data using different calibration methods: a tutorial & comparison of methods

by

Wynand-Junior Van Staden

Thesis presented in partial fulfilment of the requirements for the degree of Master of Science in Mathematics in the Faculty of Science at Stellenbosch University

Department of Mathematical Sciences,
University of Stellenbosch,
Private Bag X1, Matieland 7602, South Africa.

Supervisor:

Dr. M. Hazelbag

Co-supervisor:
Prof. W. Delva

March 2020

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Signature: .....
W. van Staden

Date: ..... March 2020

Abstract

**Calibrating a stochastic SIR-model to simulated data using different calibration methods:
a tutorial & comparison of methods**

W. van Staden

Department of Mathematical Sciences,
University of Stellenbosch,
Private Bag X1, Matieland 7602, South Africa.

Thesis: MSc. (Mathematics)

March 2020

Mathematical models have helped researchers identify and quantify trends in observed data, which is especially useful in the field of epidemiology. Fitting models to data enhances the credibility of model results, since the underlying framework of disease, is quantified and epidemiological drivers can be found. However, many calibration methods exist that quantify key parameters of a model, given observed data, and choosing which calibration method to use in a study needs justification. Also, understanding how different calibration methods work, can improve the quality and reduce uncertainty of estimated parameters. Four calibration methods (two optimization methods and two sampling methods) were reviewed and compared by calibrating a simple stochastic SIR model to model simulated data, with all four methods. With the target parameters known and by evaluating the performance of the calibration methods by using bias, accuracy and coverage measures, it was found that sampling methods (Bayesian Maximum Likelihood Estimation and the Approximate Bayesian Computation rejection algorithm) outperform optimization methods (Least Squares and Maximum Likelihood Estimation).

.

Keywords: Calibration methods, parameter estimation, simulation study, SIR model.

Opsomming

Kalibrering van 'n stogastiese SIR-model na gesimuleerde data met behulp van verskillende kalibrasiemetodes: 'n tutoriaal & vergelyking van metodes

("Calibrating a stochastic SIR-model to simulated data using different calibration methods: a tutorial & comparison of methods ")

W. van Staden

Departement Wiskundige Wetenskappe,
Universiteit van Stellenbosch,
Privaatsak X1, Matieland 7602, Suid Afrika.

Tesis: MSc. (Wiskunde)

Maart 2020

Wiskundige modelle help navorses om die neigings in waargeneemde data te identifiseer en te kwantifiseer, wat veral nuttig in die van epidemiologie is. Deur modelle aan data te kalibreer, word die geloofwaardigheid van model resultate verhoog, aangesien die onderliggende raamwerk van 'n siekte gekwantifiseer word en epidemiologiese drywers gevind kan word. Daar bestaan egter baie kalibrasiemetodes wat die sleutel parameters van 'n model kwantifiseer, gegewe waargenome data en die keuse van die kalibrasiemetode om in 'n studie te gebruik, moet gereverdig word. Deur om te verstaan hoe verskillende kalibrasiemetodes werk, kan dit die kwaliteit verbeter en onsekerheid van geskatte parameters verminder. Vier kalibrasiemetodes (twee optimeringsmetodes en twee steekproef metodes) is hersien en vergelyk deur 'n eenvoudige stogastiese SIR-model te kalibreer aan gesimuleerde data met al vier metodes te modelleer. Met die teikenparameters bekend en deur die werking van die kalibrasiemetodes te evalueer deur die berekening van vooroordeligheid, akkuraatheid en bedekking, is daar gevind dat steekproefmetodes (Bayesian Maximum Likelihood Estimation en die Approximate Bayesian Computation verwerpings algoritme) beter as optimeringsme-

todes (Least Squares en Maximum Likelihood Estimation) vaar.

Sleutelwoorde: Kalibrasiemetodes, parameter skatting, simulerings studie, SIR model.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Dr. Marijn Hazelbag, co-supervisor, Prof. Wim Delva and training director, Masimba Paradza, who have helped me grow as a researcher and an independent thinker and worker as well as giving so much insight to this work.
Thank you to the DST-NRF Centre of Excellence in Epidemiological Modelling and Analysis (SACEMA) for funding me in the duration of this study. Also, thank you to SACEMA for granting me entry to so many courses that has improved my research, analytic and quantitative abilities during my time at SACEMA.
I would like to thank all my family and friends for all their support and prayers. I would also like to thank all my colleagues at SACEMA for all the quick and meaningful conversations as well as the support.
Above all, I would like to thank God for granting me the strength and determination to have been able to complete this work.

Dedications

To Lynn-Lee Ann and Rio Jesse.
Also my mom, dad and sister.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

1.1.1 Calibration of models to data

Mathematical modeling allows researchers to identify and quantify trends in observed data from which future trends can be predicted/generated, which is especially useful in the field of epidemiology. Trends in disease data need to be studied to help prevent epidemics and in some cases improve treatment strategies. Mathematical models can be designed to follow the real-world dynamics of diseases by quantifying transmission and subsequent recovery from diseases within a population, as seen in compartmental models. When models are fitted to the observed disease data of a population, researchers have a quantitative framework with which the underlying mechanisms of a disease can be studied (Chowell, 2017). This gives researchers the tools to make predictions of future disease prevalence and assess how interventions may influence disease incidence. By being able to simulate epidemic trajectories by exploring different scenarios using models, insights into the key epidemiological drivers of diseases are found (Punyacharoensin *et al.*, 2011). Thus by fitting models to observed data researchers can draw more reliable conclusions, based on true disease dynamics within populations. Also, when model inputs, i.e. model parameter values, are informed by observed data, the credibility of the model results is enhanced. (Taylor *et al.*, 2010).

The process of fitting models to data and finding the values of the model input parameters that may have generated the data is called calibration. Calibration is a procedure that adjusts unknown parameter values by comparing various outputs of data generated by a model (using different parameter values) to observed data (Vanni *et al.*, 2011).

Calibration is a very important step if no information from previous research is available to inform the value of certain parameters. These parameters are then calibrated using a calibration method, a mathematical model and observed data. Calibration is thus a critical step to establish credibility in modeling (Stout *et al.*, 2009).
Calibration methods depend on a few key components to establish this credibility:

- Parameter search strategy,
- G.o.F measure,
- Acceptance criteria and
- A stopping rule.

These components are briefly described in the following table 1.1:

Table 1.1: Key components of calibration methods

Component	**Description**
Target statistics	This refers to the summary statistics of the observed data that the calibration method attempts to replicate. The researcher chooses the target statistics that are key to the data and holds the most value in estimating parameters. The summary statistic may be a single statistic (age of a patient, etc.) or a series of statistics (certain time points in a data curve) (Stout *et al.*, 2009)
Parameter search strategy	The core algorithm a calibration method uses to locate and estimate parameters. Calibrations methods can be divided into two groups, optimization, and sampling methods, depending on its parameter search algorithm. Optimization methods use optimized path algorithms to search for parameters whereas sampling methods depend on drawing from prior distributions to search for parameters.

Component	Description
G.o.F	A calculation/metric that is used to compare the target summary statistics of the observed data to the same summary statistics of the model output data produced by explored input parameters. The G.o.F measure is used in the parameter search algorithm, which ultimately quantifies how accurate and valid the explored parameters are by the calculation of the calibration method's specific objective function.
Acceptance criteria	The criteria to be met for parameters to be accepted as the estimated parameters. The acceptance criteria compares the results from the G.o.F measure to a threshold value. When parameters produce model output data that allows the G.o.F measure to meet the threshold criteria/requirements, the input parameters are accepted by the calibration method. The acceptance criteria are thus rules for defining which parameter combinations will be included in the output of the calibration.
Stopping rule	The stopping rule defines when the parameter search can stop. This can be a tiered rule i.e. stop as soon as either of the following two things is true: • parameter combinations have been accepted under the acceptance criteria, • a specified number runs have taken place.

1.1.2 Methods of model calibration

Calibration is the process of determining key parameters of a model fit to data but calibration methods vary from estimation methods since estimation methods do not evaluate the overall fit of a model to observed data (Stout *et al.*, 2009). Most literature agrees that calibration is needed to maintain the credibility of model prediction, thus understanding the key components of the calibration method that will be used in a study is a crucial step. Many calibration methods exist and they all aim to either minimize or maximize their respective objective function when comparing model output data to the

observed data of the study (Taylor *et al.*, 2010). Since the calibration method focuses on model output data to estimate parameters, the key components of the calibration method have to be tailored to the needs of the model used in the study (Dahabreh *et al.*, 2017).

The researcher needs to assess how identifiable the model parameters are and what impact the choice of calibration targets (summary statistics) of the observed data has on the estimation of the model parameters by the calibration method (Dahabreh *et al.*, 2017). When the model used in the study is well identified, the researcher needs to systematically asses how a calibration method can effectively estimate the unknown model parameters values. In choosing a calibration method, the key components of the calibration method need to be assessed by posing the following questions (Dahabreh *et al.*, 2017):

1. How well does the parameter search strategy optimize the domain of the objective function of the G.o.F measure?
2. How accurate does the objective function of the G.o.F measure quantify the model fit?
3. How well does the acceptance criteria define convergence?
4. How well does the stopping rule define an exhaustive parameter search?

By answering these questions the choice and implementation of a certain calibration method can be justified. If previous studies give conflicting values to parameters or there is a lack of empirical data to inform the choice of parameter values used in a model, having a calibration method that decreases parameter uncertainty then also gives more validity to the estimated parameter values (Briggs *et al.*, 2012). Having confidence in key components of a calibration method emphasizes the importance of the relationship between estimated parameter values and the credibility of model output (Briggs *et al.*, 2012).

There are many different types of calibration methods and some overlap in their defined key components. The researcher, however, needs to identify how well these key components of the respective calibration methods would address the parameter uncertainty and estimation in their study.

Here, the calibration methods have been grouped according to the type of parameter search strategy employed and the type of output the calibration method returns:

- Optimization methods
- Sampling methods.

1.1.2.1 Optimization methods

Optimization methods are methods that employ a parameter search strategy that uses an optimized path to search and locate feasible parameters. These methods incorporate its G.o.F measure in its movement through the parameter space, by evaluating the objective function at every parameter or parameter combinations in the parameter space. Once the acceptance criteria or subsequent stopping rule is adhered to, the method returns the parameters that allowed for the objective function to meet the threshold criteria. Results from optimization methods suit the Frequentist approach, where the standard errors of the estimated parameters can be calculated, from which confidence intervals may be derived. Optimization methods are mostly differentiated from another through their respective algorithm employed by the parameter search strategy and G.o.F measure.

Two very popular optimization methods are the Least-Squares (LS) and Maximum Likelihood Estimation (MLE) methods. In this study, these methods are known as calibration methods, since they can define distinct G.o.F measures and can be implemented using an optimized parameter search algorithm.

Least-Squares
The Least-Squares (LS) method is a calibration method that minimizes the squared distances from observed data to data produced by a model using explored parameters (Van De Geer, 2005). LS estimation was first published by Legendre in 1806, but historians believe the method was first developed by Gauss in 1795 (Sorenson, 1970). He addressed some important points with the use of the LS method (Sorenson, 1970):

1. The number of suitable observations (summary statistics of the observed data) is very important for the determination of unknown parameter values.
2. The residuals, i.e. the difference between the observed data and model output data using the explored parameters, have to be very small so that the parameter estimates may very accurately replicate the observed data.

3. The inaccuracies of recording the observed data may lead to better parameter estimation using probabilistic techniques.

The LS method aims to find the parameter values that give the minimum sum of the squared residuals, i.e. the difference in what is seen in the observed data and the model output data produced by the explored parameters. Guass named it the most probable value of the parameter since it produces the smallest sum of squared residuals (Sorenson, 1970).
Finding the best-suited parameter value using the LS method thus becomes a matter of finding the parameter x_p that produces the smallest y_p using the following formula:

$$y_p = \sum_{i=1}^{n} \left(MO_i(x_p) - TS_i\right)^2,$$

where n is the number of summary/target statistics, $MO_i(x_p)$ is the model output summary statistics using the explored parameter x_p and TS_i is the target statistics of the observed data. The x_p value that corresponds to the smallest y_p value is then returned as the parameter estimate $\hat{x}_p$, which produced the least sum of squares of its residuals. By using the LS method, a confidence interval (CI) can be constructed for the parameter estimate by using the variance (Van De Geer, 2005):

$$CI_p = \hat{x}_p \pm c\sqrt{var(\hat{x}_p)},$$

where changing the value of c gives a different CI range, i.e. $c = 1.96$ gives the 95% CI.

Maximum Likelihood Estimation
The Maximum Likelihood Estimation (MLE) method is a calibration method that seeks the parameters that maximize the likelihood function and ultimately are the most likely to have produced the observed data (Myung, 2003). The method was developed and improved by R.A. Fisher between 1912 when he first presented the numerical procedure and 1922 (Aldrich, 1997). To find the parameter estimate that is the most likely of being the parameter that produced the observed data, the product of the likelihood functions of the individual observations (given the observations are independent) need to be found (Myung, 2003). The likelihood function $L(x_p|TS)$ of a parameter x_p, is thus the product of the probability density functions of the model output summary statistics produced by the parameter, $MO(x_p)$, given the target statistics of the observed data TS and the number of observations in the observed data (i.e. sample population size) N:

$$L(x_p|TS) = \prod_{i=1}^{n} \binom{N}{TS_i} MO_i(x_p)^{TS_i}(1 - MO_i(x_p))^{N-TS_i},$$

where n is the number of summary/target statistics.
After finding the likelihood function values for all the explored parameter (i.e. p = [1, m]), the MLE method now consists of finding the parameter that produced the maximum likelihood function value. The most likely parameter is then found by maximizing the log-likelihood function $log(L(x_p|TS))$, where the log function is the natural logarithm (ln). The parameter estimate $\hat{x}_p$ found by using the MLE method is thus deemed to have the most likely probability distribution given the observed data (Myung, 2003).

To ensure that the MLE method finds the maximum parameter estimate the function $log(L(x_p|TS))$ has to be differentiable, so that

$$\frac{\partial log(L(x_p|TS))}{\partial x_P} = 0,$$
$$\frac{\partial^2 log(L(x_p|TS))}{\partial x_P^2} < 0,$$

thus when the partial derivative is 0, local maximum or minimum is found and with the second partial derivative being negative ensures that a local maximum is found. However, the problem does exist that the local maximum found, is not the global maximum of the log-likelihood function. Depending on the optimization algorithm, when the parameter search starting value is closer to a certain local maximum, the MLE method might not converge to the global maximum of the log-likelihood function (Myung, 2003). This issue is usually solved by doing many iterations of the calibration, with multiple starting values.

1.1.2.2 Sampling methods

Sampling methods are methods that focus on prior knowledge of the model parameters and require informed prior distributions. These methods randomly draw parameters from the informed prior distributions of model parameters and compares the respective model output produced using the prior distributions to observed data. The set of parameters that are the most feasible (feasible parameters depends on the acceptance criteria of the calibration method) are then stored in a posterior distribution of parameters. Sampling methods are based on a Bayesian approach, where prior knowledge of the model parameters have a big influence on the outcome of parameter estimation.

Two popular sampling methods are Bayesian inference (this study will look at Bayesian Maximum Likelihood Estimation) and Approximate Bayesian Computation. Bayesian

methods are popular since they can synthesize model parameters based on model outcomes and parameter weights. These methods aim to find the well-fitting sets of parameters based on evidence provided by comparing model output data to the observed data (Menzies *et al.*, 2017)

Bayesian Maximum Likelihood Estimation

Bayesian inference was named after T. Bayes who had a paper published posthumously in 1763, that described a specific example. The Bayesian interpretation of probabilities was then numerically developed by Laplace between the late 1700s and early 1800s, with the history of Bayesian inference described in Fienberg (2006)

Bayesian inference calibration methods follow a quantitative approach to finding parameter sets that are the most plausible in producing the observed data (Jackson *et al.*, 2015). This approach involves:

1. Defining plausible ranges for prior distributions to draw parameters from.
2. Comparing the model output from the drawn parameters to the observed data.
3. Placing weights on the parameters given the G.o.F measure.
4. Retaining a subset of the parameters that are the most plausible, given the acceptance criteria of the method.

Bayesian Maximum Likelihood Estimation (BMLE) uses this approach with its G.o.F measure being the likelihood function, as described by the MLE method. The BMLE method takes into account the evidence-based on the prior probability distributions $p(\theta)$ and the likelihood function $p(Y|\theta)$ of the observed data given the parameters in the prior distribution (Menzies *et al.*, 2017). The BMLE method then attempts to find the posterior probability distribution $p(\theta|Y)$ by:

$$p(\theta|Y) \propto p(\theta) \times p(Y|\theta),$$

where some applications of the method scales the posterior distribution by multiplying by $\frac{1}{p(Y)}$, with $p(Y)$ being the probability of observing the data.

The BMLE method uses a sampling importance re-sampling (SIRS) method which is an algorithm that determines how parameters are selected to be stored in the posterior distribution. The SIRS algorithm involves finding model output from the parameters in the prior distribution, finding the likelihood function value for each of the parameters and selecting the subset of parameters with the highest likelihood function values by

re-sampling parameters from the prior distribution according to the weights (given by the likelihood function values) (Menzies *et al.*, 2017).

Approximate Bayesian Computation

The Approximate Bayesian Computation (ABC) methods are Bayesian inference methods that do not rely on likelihood approximations to find the most plausible parameter estimates (Wilkinson, 2008), with the ABC rejection algorithm (ABC-r) being the most simple version. Ideas around the ABC methods were first introduced by Diggle & Grattle in 1984 and also Rubin in 1984 (Beaumont, 2010), but the methods were formally established and proposed by Beaumont *et al.* (2002).

For the ABC-r method it is assumed that all the parameters have an independent prior distribution $p(\theta)$, from which a large number of parameters are drawn and model output ($MO(\theta)$) is produced from each of the sampled parameters (van der Vaart *et al.*, 2015). The model output summary statistics are then compared to the target statistics of the observed data using a distance measure $d(MO(\theta), TS)$ (usually specified in a study). A subset of a specified tolerance amount of parameters (δ) that produced the smallest distances are then accepted into a posterior distribution $p(\theta|MO(\theta))$. The ABC-r method is thus a three-step calibration method (Wilkinson, 2008):

1. Draw θ from $p(\theta)$.
2. Generate model output $MO(\theta)$.
3. Accept θ if $d(MO(\theta), TS) \leq \delta$.

Smaller values of δ would thus lead to better posterior distribution approximates of parameters, however, this would also lead to fewer values being accepted, thus more computations will be needed to find large enough posterior distributions for further analysis (Wilkinson, 2008). Thus there is a trade-off between computational efficiency and accuracy in the ABC-r method. The ABC-r method thus differs from Bayesian inference methods since it provides a systematic way for parameters to be estimated based on the support that different models and observed data provide different studies (van der Vaart *et al.*, 2015).

1.1.3 Numerical examples of calibration methods

1.1.3.1 Least-Squares

The Van De Geer (2005) study performed a linear regression and used the LS method to estimate the coefficients of the regression function. The general form of the regression function was given as

$$f_\beta(X) = \beta_1 + X\beta_2 + X^2\beta_3,$$

and the LS estimator is expressed as $\|\mathbf{y} - \mathbf{X}b\|^2$, where $\mathbf{y}$ is a vector of the response variables, $\mathbf{X}$ is a 100×3 data matrix, containing 100 observations of 3 parameters and b is the explored parameters. The LS estimator (to find parameter estimates $\hat{\beta}$) was then given by

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}\mathbf{y}.$$

The target parameters were $(\beta_1, \beta_2, \beta_3) = (1, -3, 0)$ and the parameter estimates found using the LS estimation were $(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3) = (0.5778, -2.3856, -0.0446)$. The study then went on to just calculate the confidence intervals for the $\hat{\beta}_3$ parameter:

$$\begin{aligned} CI_{\hat{\beta}_3} &= \hat{\beta}_3 \pm c\sqrt{var(\hat{x}_p)} \\ &= -0.0446 \pm 1.96\sqrt{1.5902} \\ &= [-2.5162, 2.470] \end{aligned}$$

The study went on to conclude that given the confidence intervals, the null hypothesis that $\hat{\beta}_3 = 0$ can not be rejected and then estimated the parameters β_1 and β_2 again, finding

$$(\hat{\beta}_{1,0}, \hat{\beta}_{2,0}) = (0.5854, -2.4306) \neq (\hat{\beta}_1, \hat{\beta}_2).$$

Thus giving a conclusion that changing the number of parameters to estimate results in reducing the difference between the target parameters and the parameter estimates from the LS method.

1.1.3.2 Maximum Likelihood Estimation

The numerical example for the MLE method comes from the study of Beerli (2006), where 10 loci datasets were analyzed that contained 20 individuals with data from four

groups of 100 single locus populations. The effective population size (Θ_4) and the migration rate of populations from group 1 to 4 (M_{14}) were estimated using the MLE method and MIGRATE software, where the likelihood of the given the data model parameters was computed by:

$$L(D|\pi) = \sum_T \int_B k(T, B|\pi) L(D|T, B) dB,$$

where $k(T, B|\pi)$ was reported as the Kingman coalescent probability density and $L(D|T, B)$ is the likelihood of the data given the genealogy. The calibration ran for 100 runs where the median ($\hat{\Theta}_4$ and $\hat{M}_{14}$), 25% and 75% quartiles were reported as the results as well the coverage of the 95% confidence interval of the target parameters.
The study varied the target parameter, the effective population size $\Theta_4 = (0.0001, 0.001, 0.01, 0.1)$ and kept the migration rate parameter constant $M_{14} = 100$. The results for the MLE method were reported as follows:

Table 1.2: The results of the MLE method from estimating Θ_4 and M_{14} by Beerli (2006)

Θ_4	$\hat{\Theta}$	Coverage(Θ_4)	M_{14}	$\hat{M}_{14}$	Coverage(M_{14})
0.0001	0.00092	6%	100	0.2	33%
0.001	0.0017	47%	100	46.3	55%
0.01	0.0104	94%	100	53.7	62%
0.1	0.0573	51%	100	66.5	49%

The study concluded that using the MLE method with data that does not contain much information results in inadequate convergence. As seen in the results of the study, the MLE method did not estimate M_{14} well, especially when $\Theta_4 = 0.0001$. It can be noted however that as the effective population size increased the estimation of both parameters improved, expect when Θ_4 increased from 0.01 to 0.1.

1.1.3.3 Bayesian Maximum Likelihood Estimation

The Menzies *et al.* (2017) study used and described the BMLE method to inform policy on health burden, budget impact, and cost-effectiveness. A compartmental model was designed to simulate data for a hypothetical disease where the analysis centered around estimating the cost-effectiveness of treatment against the stage of the disease progression.

The model consisted of the following health state variables:

- N - non-susceptible,
- S - susceptible,
- E - early disease,
- L - late disease,
- T - treatment,
- D - dead.

Data of a period of 30 years were simulated and 20 years of the data were used in the analysis, in which the incremental cost-effectiveness ratio (ICER) was calculated, which was defined as the ratio of incremental cost to incremental life-years lived for the proposed policy. The model contained 11 parameters (see Menzies *et al.* (2017) and table 1.3) of which only 7 were calibrated to the simulated data since 4 of the parameters were not relevant to the ICER obtained. The study compared the results from running the model with all the 11 parameters without calibration and the results from running the model with the 7 calibrated parameters. In the calibration however the SIRS algorithm was improved upon because from 100000 parameter draws, the re-sampled posterior distribution only contained 797 unique parameter sets. This resulted in a low effective sample size of 88, whereas the incremental-mixture importance algorithm (IMIS, unspecified in the study) produced a re-sample posterior of 10000 parameters, with 6372 unique parameter sets.
The parameters, prior distributions and calibrated posterior distributions are reported in table 1.3:

Table 1.3: Model parameters with prior and posterior distributions from the calibration

Parameter	**Description**	**Prior distribution (Mean (95% CR interval))**	**Posterior distribution (Mean (95% CR interval))**
b	Fraction of births entering non-susceptible state	0.200 (0.03, 0.48)	0.212 (0.17, 0.26)
μ^E	Disease-specific mortality for early disease	0.050 (0.02, 0.12)	0.040 (0.02, 0.08)
μ^L	Disease-specific mortality for late disease	0.250 (0.08, 0.59)	0.165 (0.09, 0.29)
μ^T	Disease-specific mortality on treatment	0.025 (0.01, 0.06)	0.022 (0.01, 0.04)
ρ	Effective contact rate for transmission	0.500 (0.17, 1.18)	0.540 (0.49, 0.60)
p	Rate of progression from early to late disease	0.100 (0.03, 0.24)	0.131 (0.08, 0.21)
r^L	Rate of treatment uptake for late disease	0.500 (0.17, 1.18)	0.585 (0.24, 1.24)

The uncalibrated model found an ICER of US $1300 per life-year saved, whereas the calibrated model found a lower ICER of US $947 per life-year saved. It can also be seen that the credible intervals (CrI) of all the posterior distributions of the parameters are more narrow than the prior distribution CrI's.

The study concluded that the choice of priors, likelihoods and the model for complicated policy decision making is a difficult task, even if proven evidence and guidance are found in the literature.

1.1.3.4 Approximate Bayesian Computation rejection algorithm

The van der Vaart *et al.* (2015) study used the ABC-r method in an ecological modeling study where an individual-based model (IBM) with 14 parameters that describe the dynamic energy budgets of individual earthworms. The earthworm data were simulated by Johnston *et al.* (2014) which describes the energy consumption and food uptake of earthworm populations. The van der Vaart *et al.* (2015) study implemented the ABC-r method by running simulations parallel on ARCHER software and further analysis were done in R using the `RNetLogo` package, which allowed for the use of NetLogo within R. The data was simulated with two models, a full model and a simplified model, where the parameters were derived from the Johnston *et al.* (2014) study as described in table 1.4:

Table 1.4: Model parameters for the simulation of data (mean values of log-normal prior distributions)

Parameter	Description	Parameter values (mean)
B_0	Taxon-specific normalization constant	967
E	Activation energy	0.25
E_c	Energy cost of tissue	3.6
E_f	Energy from food	10.6
E_s	Energy cost from synthesis	3.6
h	Half saturation coefficient	3.5
IG_m	Maximum ingestion rate	0.70
M_b	Mass at birth	0.011
M_c	Mass of cocoon	0.015
M_m	Maximum asymptotic weight	0.5
M_p	Mass at sexual maturity	0.25
r_B	Growth constant	0.177
r_m	Maximum energy to reproduction	0.182
s	Movement speed	0.004

However in the simplified model the IG_m parameter value was changed to 0.15. 1000000 simulations were run and 100 of the best fitting parameters were selected based on the following distance measure:

$$\rho(m_i, D) = \sqrt{\sum_j \left(\frac{m_{i,j} - D_j}{sd(m_j)} \right)},$$

where $m_{i,j}$ is run i's output for data point j, D_j is the empirical data for data point j and $sd(m_j)$ is the standard deviation for data point j from all the model runs. By dividing by $sd(m_j)$ the parameters were scaled and normalized, which allowed for better comparison of results between parameters. The calibration was implemented once for the full model and then again for the full model and the simplified model. The fit of the parameters were then evaluated by

$$R^2 = 1 - \frac{\sum_j (m_j - D_j)^2}{\sum_j (D_j - \overline{D})^2},$$

where R^2 is the proportion of the variance for each experiment and $\overline{D}$ is the mean of the empirical data in that experiment. The study measured the results of the ABC-r method in four different ways: comparing the prior distribution to the posterior; comparing the

R^2 values the ABC-r method to that of running the IBM 100 times; cross-validation by setting aside 100 random model outputs, performing ABC-r on the remaining runs and comparing the medians of the accepted parameters with the target parameter values; and coverage by taking the 100 best runs as pseudo-data, using the ABC-r method on the remaining runs and then calculating the relative frequency of the accepted parameter values being less than that of the pseudo data.

The marginal posterior distributions for seven of the parameters (E, IG_m, M_b, M_m, M_p, r_B, r_m) were narrower than the corresponding priors. In 3 out of 6 experiments, the ABC-r method had produced better R^2 values than the IBM model run 100 times. From the cross validation, seven of the parameters (M_b, M_m, r_B, E, IG_m, M_p, r_m) had narrowed posteriors and were strongly correlated with the target parameter value. From the coverage evaluation, it was found that nine parameters (B_0, E, E_c, E_f, E_s, M_c, M_p, r_m, s) had uniform posterior distributions and it was reported that the coverage has held.

The study further concluded that from these results the ABC-r method provided slightly better fits than that of literature and that the method is able to facilitate complexities in model selection, parameterization and uncertainty analysis.

1.2 Problem statement

All of the calibration methods that are described here have such unique ways of estimating parameters by calibrating models to observed data, that it makes it hard for a researcher (uninformed one) to choose which method is best to use for their specific study. When it is a necessary step in a study to calibrate a model to observed data to find the best fitting parameters, especially when values for these parameters are not found in literature, the choice of a calibration method becomes an integral part of the study which needs as much justification as incorporating parameter values from previous studies.
Also, given that models have to make assumptions regarding real-world phenomena and potential data inconsistency (missing data or outliers) being able to find parameter estimates and quantify uncertainties i.e standard errors, gives more credibility to the results found in the specific study (Pernot and Calliez, 2017).

Uncertainty around parameter estimates are quantified by finding standard errors and confidence intervals (CI's), which infers the precision of the calibration method but these quantities are only found for point estimates, which are generally the results from opti-

mization methods (Briggs *et al.*, 2012). However, sampling methods provide parameters sets which are subsets of the best fitting parameters given the observed data, where the precision and quality of the posterior distributions depend on the specifications of the informed prior distribution. Credible intervals (CrI's) can be found for posterior distribution without the use of standard errors, but how accurate would the coverage of a CrI be in comparison with a CI?

The problem is that there aren't many studies that compare a wide range of calibration methods to each other, especially well-established optimization methods and well-established sampling methods. Also, there aren't many studies that use simple models, identifiable target statistics and target parameters, to review individual calibration methods or to compare calibration methods.

Dahabreh *et al.* (2017) mentions four studies that either compared alternative calibration methods or different key components to the same problem:

- Kong *et al.* (2009): Compared a simulated annealing algorithm (optimization method - initial parameter combinations are randomly selected and G.o.F measure is calculated) to a genetic algorithm (sampling method - G.o.F measure is calculated for every parameter). The calibration methods were compared by accuracy - which method could produce the lowest G.o.F measure, and speed - which method could reach the specified stopping rule first.
- Taylor *et al.* (2010): Compared a random search algorithm (sampling method - randomly drawing 100000 parameters from a prior distribution), a manual calibration (an analyst that manually adjusts parameters) and the Nelder-Mead algorithm (optimization method). The calibration methods were compared by finding the parameter set that minimizes a specified weighted mean deviation G.o.F measure (the random search algorithm also only returned a parameter set and not a posterior distribution of parameters).
- Karnon and Vanni (2011): Compared a random search algorithm (sampling method - randomly draws parameters from prior and 1000 best-fitting parameters were returned) and a generalized reduced gradient method (optimization method - moves along a gradient from a starting point to locate a minimum point). The study also compared two G.o.F metrics, chi-squared and likelihood, using both calibration methods as well as different convergence (acceptance) criteria for the random search algorithm.

- Taylor *et al.* (2012): Compared five different random starting values for the calibration of a Markov cohort model using the Nelder-Mead method algorithm (optimization method).

Given that Dahabreh *et al.* (2017) study was conducted in 2017, indicate that not many studies have compared and reviewed the performance of calibration methods. The sampling methods that were used in these reported were not as well-established as the standard Bayesian inference methods, especially the BMLE method, nor the ABC methods. The Kong *et al.* (2009), Taylor *et al.* (2010) and Taylor *et al.* (2012) studies also did not give complete specifications of the models that were used to compare their respective calibration.

This has motivated the work put forward in this study, to compare well-established calibration methods of different types, with different key components, by using a common mathematical model. This study has to provide a framework in which calibration methods can easily be reviewed and the performance of different types of methods can easily be compared.

1.3 Study objectives

There is a lack of attention going into comparing different calibration methods to each. There is also a lack of meaningful study design and framework to compare different calibration methods, especially comparing methods that return different types of results from respective calibrations (i.e. parameter estimates vs posterior distributions of parameters).

This study attempted to provide a simple framework to compare different types of calibration methods, as well as give a tutorial on how the optimization methods, LS and MLE, and sampling methods, BMLE and ABC-r can be used to calibrate a simple model to data.

1.3.1 Research Question

Given the motivation to address the problem of comparing different calibration methods, the following research questions were formulated:

1. How well can different calibration methods, in the same scenarios:

a) Minimize bias?

b) Maximize accuracy?

c) Find sufficient coverage of the target parameters?

2. How does the performance of the calibration methods change according to:

a) Number of target statistics?

b) Sample size?

c) Number of parameters to estimate?

3. Which calibration method performed the overall best in this study?

1.3.2 Study implementation

This study was implemented using a simulation study. Simulation studies are studies that can test the accuracy and performance of different statistical methods using computer-intensive techniques (Burton *et al.*, 2006). Implementing a simulation study enables for multiple calibrations to be run so that multiple parameter estimates can be found. This allows for performance measuring of the calibration method since the truth about all the input data is known as (Burton *et al.*, 2006).

With the implementation of a simulation study, Bias, Accuracy and Coverage values of the results from the calibration methods were calculated to evaluate the performance of the calibration methods, which was then used to compare the methods to each other. It is good practice to implement different performance measures, to be able to validate the precision of the calibration methods, since results may vary per measure (Burton *et al.*, 2006).

Also, as seen in results from studies in the numerical examples above, estimating parameters using different target data, population sample size and number of parameters to estimate can impact the results of calibration methods. Thus the impact of changing these variables was also studied here.
Ultimately the study also aims to evaluate which of the four calibration methods has the best performance given certain conditions.

Chapter 2

Methods

2.1 Introduction

This chapter aims to give clarity on how this study was conducted and the reasoning behind the design choices. The research design section in this chapter is organized as follows, by describing the use of a simulation study; describing the mathematical model used to simulate data and calibrate parameters; describing the scope of the explored scenarios in the study; describing the methodology behind finding results and how to compare the calibration methods; describing how the calibration methods were implemented in this study; and then describing the performance measures that were used to measure the performance of the different calibration methods.

2.2 Research Design

2.2.1 Simulation Study

A simulation study was conducted using R and R Studio software. Simulations studies are particularly useful types of studies when it comes to understanding concepts of statistical methods. The assessment of certain aspects of statistical methods may require a study design in which the method is applied a high number of times (hundreds or thousands of repetitions) in a controlled manner.
In using a simulation study design in this study allowed the generation of many simulated data sets which ultimately allowed for rigorous testing of the different calibration methods. Also, by implementing a simulation study, multiple calibration attempts were made possible, which allowed for better performance measuring.
Since R and R studio were used, a link to a GitHub repository containing the R code is

available in the appendix to reproduce the findings of this study or to further expand the research and results from this study.

2.2.2 Mathematical Model

A stochastic compartmental model was used during the study. The compartmental model was a simple stochastic SIR-model, which had an S-compartment (S-class) that represents susceptible individuals, an I-compartment (I-class) that represents infected individuals and an R-compartment (R-class) that represents recovered individuals. The model was implemented by using the `SimInf` package in R (background: (Widgren *et al.*, 2016), technical framework: (Widgren *et al.*, 2019)). The use of a SIR-model (especially the one implemented using the *SimInf* package in R) was motivated by the inclusion of a few properties:

- stochasticity,
- non-linear and dynamic over time, i.e. internal dependencies and feedback,
- fast execution,
- low number of parameters.

The requirements for the model was to have a model that is of the type that is commonly used in infectious epidemiology. Also, that the model was still simple enough so that the differences between the calibration methods could be explored, without getting lost in the details of the model or having to wait too long for the model runs to be completed.

The model allows individuals to flow out from one compartment to the next by a certain rate. Commonly in the SIR-model individuals move from the S-class to the I-class at a transmission rate constant (β) and from the I-class to the R-class at a recovery rate constant (γ) (as illustrated in figure 2.1).
In the model, individuals from the S- and I-classes have to make contact for individuals to move from the S- to the I-class, which is mathematically denoted as the product of the transmission rate constant (β), the amount of individuals in the S-class and the amount of individuals in the I-class divided by the total number of individuals (N), at the time step of the contact. In the limit of infinitely small steps, this can be interpreted as the probability that a susceptible individual comes in contact with an infected individual times the probability that a successful transmission occurred during the contact $\left(\frac{\beta SI}{N}\right)$.

Individuals move from the I-class to the R-class at a recovery rate constant γ times the amount of infected individuals, which could also be interpreted as: in the limit of infinitely small steps, the probability that an infectious individual successfully recovers from the disease. This is mathematically denoted as the number of infectious individuals times the recovery rate constant (γI).

Mathematically the SIR model takes the form of a set of 3 Ordinary Differential Equations (ODEs):

$$\begin{aligned}\frac{dS}{dt} &= -\frac{\beta SI}{N} \\ \frac{dI}{dt} &= \frac{\beta SI}{N} - \gamma I \\ \frac{dR}{dt} &= \gamma I,\end{aligned}$$

where $N = S + I + R$ (the total population).
The SIR-model can be expanded to include any real-world phenomena such as a birth/inflow rate and a mortality rate/outflow rate. It can also have more classes to include more types of population statuses like an exposed-class, vaccinated-class, etc. However, since the focus of this study was to evaluate the estimation of parameters, the basic version of the SIR-model was sufficient enough. This allowed for the calibration methods to only estimate the two parameters β and γ.

Figure 2.1: Flow chart of the simple SIR compartmental model

The `SimInf` package, which allowed the use of the stochastic SIR compartmental model, in R, runs at discrete-time events (instead of continuous-time) and consists of continues-time Monte Carlo Markov chains as a general model of the dynamics (the random number generator step) (Widgren *et al.*, 2016). The package incorporates the Gillespie stochastic simulation algorithm for the model's stochasticity, the same as the `GillespieSSA` package in R (Pineda-Krch, 2008), however, the `SimInf` package implements the algorithm using C code for more computational efficiency (Widgren *et al.*, 2016).

The Gillespie algorithm was initially designed to numerically simulate and quantify

the random collisions between molecules in a chemical reaction with the use of ODE's that describe the chemical reactions mathematically. The algorithm takes into account that chemical collisions occur randomly and that rate constants are more properly characterized as reaction probabilities per time unit (Gillespie, 1977). This theory is then translated to suit an epidemiological model. The steps of the algorithm are then as follows:

1. Initialization: initialize the initial state of the system (initial population conditions, rate constants, and random number generators)

2. Monte Carlo Step: generates random numbers to determine the time to next event and to determine which event occurs (either individuals move from the S- to the I-class or from the I- to the R-class).

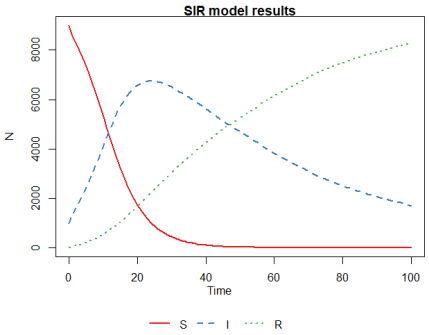
3. Update: set the time point to the point generated in Step 2 and update the number of individuals in each of the classes given the event that has occurred.

4. Iterate: repeat Step 2 and 3 until the time point is the final time point as initially specified.

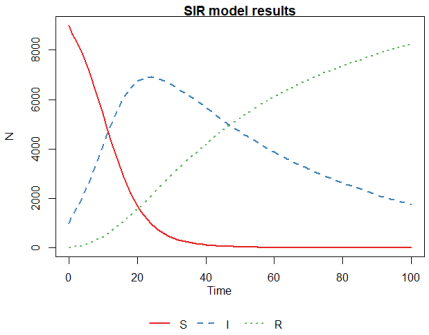
When simulating data using the SimInf package, at each time step for the next event to occur (movement of individuals from one class to another), a transition matrix is used which describes how several individuals move from one compartment to another given the transmission rate constant β and recovery rate constant γ (Widgren *et al.*, 2016). Because of the stochastic nature of the Gillespie algorithm, the outcome of the model would vary slightly even when the same β and γ parameter values were used to generate data.

To generate data using the `SimInf` package, initial conditions need to be specified for the population (number of Susceptible, Infectious and Recovered individuals) at time 0, in the form of a `data.frame()`. The initial conditions are then added to a function, `SIR()`, along with the period, β, and γ parameter values, which then generates a model. The model is then used by a `run()` function which generates the population at each of the discrete-time events (specified period) for each of the classes of the population. The output results from the `run()` function can then be used for analysis and be visualized as seen in figure 2.2.
Also, by using the same β and γ values, the model produces similar population curve trajectories, but with slightly different values, evident in figure 2.2, where the model was run with $\beta = 0.2$ and $\gamma = 0.02$ for both plots. This model stochasticity is a good representation of the fact that if a study was conducted on the same disease in the same

population (with the same disease dynamics) the study would find similar results for how the disease progresses in a population, with slightly different values.



(a) Model results plot 1



(b) Model results plot 2

	0	1	2	3	4	5	6	7	8	9	10
S	9000	8553	8284	7991	7679	7332	6955	6554	6158	5735	5307
I	1000	1391	1620	1871	2141	2421	2748	3097	3435	3782	4128
R	0	56	96	138	180	247	297	349	407	483	565

(c) Model results data 1

	0	1	2	3	4	5	6	7	8	9	10
S	9000	8624	8386	8127	7814	7441	7073	6642	6223	5778	5330
I	1000	1333	1550	1791	2067	2392	2712	3088	3441	3826	4200
R	0	43	64	82	119	167	215	270	336	396	470

(d) Model results data 2

Figure 2.2: 2.2a and 2.2b are results plots from running SIR model with the same parameters ($\beta = 0.2, \gamma = 0.02$) twice at 100 time points. 2.2c and 2.2b are the corresponding data of the population numbers at the first 10 time points in the S- I- and R-classes for plots 2.2a and 2.2b respectively

2.2.3 Scenarios

Since the relative performance of the various calibration methods may depend on several external factors, multiple scenarios were defined under which the comparison of the different calibration methods took place.

To replicate a real-world study in where the target statistics are generated from a relatively small sample of the population, the scenarios varied in the size of the sample population: between having 10% of the total population and 100% of the total population (in a perfect world study).

Since the summary statistics of the observed data is a key component of the calibration process, the scenarios had varying numbers of target statistics. In this study the summary statistics identified in the observed data were the disease prevalence at different time points, i.e. the number of individuals in the I-class (the number of infectious individuals). The calibration methods then had to identify different numbers of target statistics at different time points of the infectious prevalence curve of the observed data. The target statistics varied between 2 target statistics at time points 50 and 65; 3 target

statistics, 2 at time points 50 and 65 and another at the time point at which the infectious prevalence peaked; 4 target statistics at time points 30, 45, 60 and 75; and 64 target statistics at the time points in the interval of [1, 64].
Lastly the number of parameters the calibration methods had to estimate was also varied. Thus scenarios were varied where the methods had to estimate only γ and where the methods had to estimate β and γ.

The scenarios were thus set up as all the combinations of all of the varying components:

- Sample Population = 10% of the Total Population; 100% of the Total Population.
- Target Statistics = 2; 3; 4; 64.
- Parameters to estimate = 1 (γ) ; 2 (β, γ).

Thus the study consisted of $2 \times 4 \times 6 = 16$ scenarios of the combinations of components. Having 4 calibration methods that had to be reviewed and compared per scenario, a total of $16 \times 4 = 64$ calibrations were run in this study Also, with 1000 model runs per calibration, a total of $64 \times 1000 = 64000$ computations were performed in this study.

In the simulations the total population (N) was constant throughout all the scenarios, with $N = 10000$. From the total population, the sample population data were sampled without replacement using the base R function `sample()`. From the sampled population the Infectious individuals, at the time points corresponding to the specified time points of the number of target statistics, were extracted and then used in the calibration process. This information was then known as the observed data during each of the calibrations for each calibration method, respectively.

The simulation of data and target statistics relevant Infectious prevalence extraction was implemented by the creation of four R functions:

- `sirModel2()`: for 2 target statistics at time points 50 and 65 of the observed data,
- `sirModelPeakPrev()`: for 3 target statistics, 2 at time points 50 and 65 and another at the time point at which the I-curve attains its peak prevalence.
- `sirModel4()`: for 4 target statistics at time points 30, 45, 60 and 75 of the observed data,
- `sirModel64()`: for 64 target statistics at time points 1 to 64 of the observed data,

For each of the models in these functions, 100 time points were produced but the function only returned the time points specified by the number of target statistics of the respective functions. The returned observed data were of the form:

$$I_t^{prev} = \frac{I_t}{TotalSamplePopulation}$$

where t was the time point of the target statistic. The Infectious prevalence was divided by the total number of the sample population and returned by the function (e.g. `sirModel2()` returned two prevalence values as two decimal values).

2.2.4 Methodology

The methodology behind the design of this study was as follows:

Step 1 - Generate the target statistics using SIR-model functions and the target parameter values, with $\gamma = 0.02$ for the one-parameter calibration scenarios and $(\beta, \gamma) = (0.2, 0.02)$ for the two-parameter calibration scenarios.

Step 2 - Use the calibration methods to estimate the target parameters.

Step 3 - Repeat steps 1 and 2, 1000 times to obtain 1000 calibration attempts (i.e. parameter estimates) of the target parameters.

Step 4 - Use performance measures to evaluate the performance of the respective calibration methods within each scenario.

Steps 1 and 2 simulated how parameters are estimated in a real-world study, by using calibration on a single collected data set. Thus by repeating these steps 1000 times, results from this study were equivalent to repeating a real-world study 1000 times, to produce 1000 parameter estimates of the observed data.

2.2.5 Calibration Methods

Since the study focused on the four calibration methods, the key components of the calibration methods are specified here.

The two optimization methods, Least-Squares (LS) and Maximum Likelihood Estimation (MLE) both made use of the `optim()` function in the `stats` package in R, with the MLE method calling the `optim()` function through the `mle()` function in the `stats4` package.

The two sampling methods Bayesian Maximum Likelihood Estimation (BMLE) and Approximate Bayesian Computation rejection (ABC-r) both depended on drawing parameters from a specified prior distribution, with the prior distribution specifications being the same for both methods.

Table 2.1 describes the key components for each of the explored methods:

Table 2.1: Components of calibration methods

Calibration method	**Parameter Search Strategy**	**G.o.F**	**Acceptance criteria**	**Stopping rule**
LS (optimization)	Nelder-Mead	Square distance	Least of the sum of squared distances	Relative convergence tolerance or maximum number of iterations
MLE (optimization)	Nelder-Mead	Likelihood approximation	Maximum of the likelihood functions	Relative convergence tolerance or maximum number of iterations
BMLE (sampling)	Sampling from uniform prior distribution	Likelihood approximation	Sampling importance re-sampling	Number of re-sampled parameters
ABC-r (sampling)	Sampling from uniform prior distribution	Euclidean Distance measure	Acceptance tolerance	Number of model simulations

2.2.5.1 Least-Squares

The LS method is an optimization method that makes use of the `optim()` function in R. The *parameter search strategy* needs to be specified in the function and in this study the Nelder-Mead algorithm was used, as described in Nelder and Mead (1965).
The Nelder-Mead algorithm was developed to optimize the search of parameters that minimizes an objective function, with the use of a simplex method. The simplex method

involves having a simplex of size $n + 1$, with n being the number of parameters to be estimated, and calculating the minimization function at each vertex of the simplex (as seen in figure 2.3, the simplex takes the form of a triangle).
The Nelder-Mead algorithm for two parameters:

1. Calculate the function at each vertex.
2. Determine the vertex with highest and lowest function values, name them H and L respectively and the middle value P. Thus $f(L) < f(P) < f(H)$.
3. Construct a line from H and through the center point of L and P.
4. The simplex is transformed with regards to four candidate points on the constructed line: C_1 - a reflection of the simplex; C_2 - an expansion of the simplex; C_3 - a low side contraction of the simplex; and C_4 - a high side contraction of the simplex.
5. Calculate and evaluate the function at C_1. If $f(C_1) < f(L)$, calculate the function at C_2. Then:
   i) If $f(C_2) < f(C_1)$, replace H with C_2. The simplex expands.
   ii) If $f(C_2) > f(C_1)$ or if $f(L) < f(C_1) < f(P)$, then replace H with C_1. The simplex is reflected.
   iii) If $f(P) < f(C_1) < f(H)$ then replace H with C_3. The simplex is contracted on its low side.
   iv) If $f(H) < f(C_1)$, then replace H with C_4. The simplex is contracted on its high side.
6. Each vertex is then renamed according to step 2.
7. Steps 1 to 6 is then repeated until a convergence criteria is met.

The function the Nelder-Mead algorithm uses is the function specified in R and called by the `optim()` function.

For the LS method, the objective function was thus its *G.o.F* measure, which was the sum of squared distances between the model output summary statistics produced by the explored parameters, and the target statistics produced by using the target parameter values. Thus the `optim()` function used the Nelder-Mead algorithm to find the

parameters that gave the least sum of squared distances. The LS function has the form of:

$$\text{LS} = \sum_{i=1}^{n} (outputStats_i - targetStats_i)^2,$$

where n is the number of target statistics, $outputStats$ is the model output summary statistics from the explored parameters and $targetStats$ is the model output target statistics using the target parameter values.

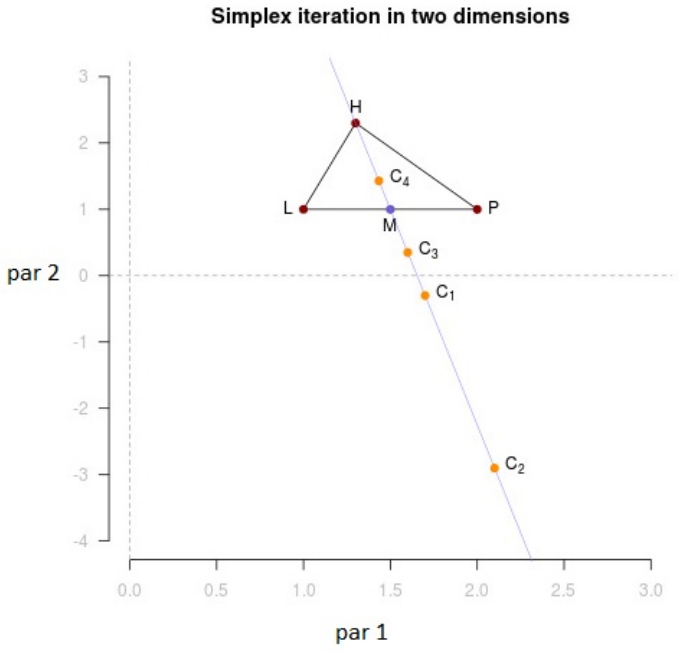


Figure 2.3: Nelder-Mead Simplex algorithm for two parameters [Dalzell (2013)]

The *acceptance criteria* for the LS method was that parameters got accepted if the objective function value of the current parameters is less than the previous objective function values of the previously explored parameters. Thus, given the conditions of the Nelder-Mead algorithm and the objective function, the parameter estimates are updated if the current parameters produce a smaller sum of squared distances. The *stopping rule* for LS method is thus when the objective function value of the current parameter estimates are less than a relative convergence tolerance value (specified by the `optim()` function in R as $1e^{-8}$) or the maximum number of 1000 iterations of the Nelder-Mead algorithm has been reached.

The starting values for the calibrations using the LS method were randomly chosen starting values using the `runif()` function in R, with the bounds of the starting value for $\beta = [0.01, 0.5]$ and $\gamma = [0.01, 0.1]$.

2.2.5.2 Maximum Likelihood Estimation

The MLE method made use of the `mle()` function which at its core uses the `optim()` function to maximize the likelihood of observing the data, given the parameter values. The Nelder-Mead method was also used as the *parameter search strategy* for the MLE method, where the maximum likelihood (ML) objective function had the form:

$$\begin{aligned} \text{ML} = -\Bigg[\sum_{i=1}^{n} log\binom{SS}{targetStats_i} + targetStats_i \times log(outputStats_i) \\ + (SS - targetStats_i) \times log(1 - outputStats_i))\Bigg] \end{aligned}$$

where n is the number of target statistics, SS is the total sampled population size, $outputData$ is the model output summary statistics produced by the explored parameters and $targetStats$ is the model output target statistics produced by the target parameter values. The logarithm used here was the natural logarithm (ln) with base e.
The *G.o.F* measure the MLE method used was the likelihood approximation of the model output data using the explored parameters given the observed data. The ML objective function had the form of a negative log-likelihood since the likelihood values produced were so small, R perceived these values as 0's, thus by taking the log of the small values resulted in log-likelihood values that were negative, with the most likely value being the closest to 0. Since the `optim()` function minimizes the objective function it makes use of, by taking the negative of these log-likelihoods allowed for the more likely parameter values (smallest negative values) to have the smaller the negative log-likelihood values (smallest positive values). Thus the Nelder-Mead algorithm still had to succeed in minimizing the ML objective function (finding the smallest objective function value) which translated as maximizing the likelihood.

The *acceptance criteria* and the *stopping rule* of the MLE method were the same as that of the LS method. As well as how starting values for the calibrations were chosen, with the bounds for the parameters the same for the MLE method as for the LS method.

2.2.5.3 Bayesian Maximum Likelihood Estimation

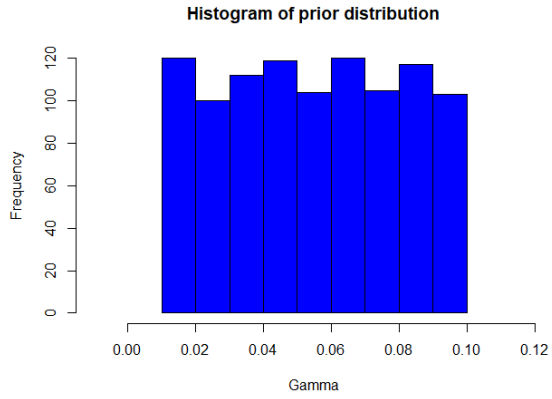
The BMLE method is a Sampling method that made use of the Sampling Importance Re-sampling *parameter search strategy* as described in [(Menzies *et al.*, 2017)]:

1. Parameters were randomly drawn from a uniform prior distribution.
2. The randomly drawn parameters were then evaluated using the same *G.o.F* measure as the MLE method (likelihood approximation):
   i) Model output summary statistics were produced using the randomly drawn parameters.
   ii) The model output summary statistics were then compared to the target statistics using the same ML function that the MLE method used, however, the negative of the log-likelihood values were not taken in the BMLE method. The resulted log-likelihood values were then assigned to each of the parameters (or parameter combinations) that were used to produce the model output summary statistics.
3. The log-likelihood values were then converted to weights by the formula:

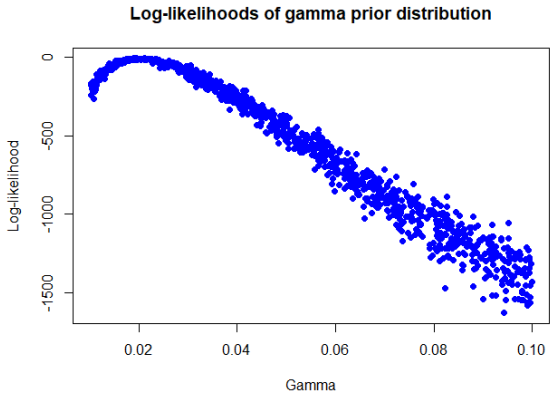
$$Weight_p = \frac{e^{ll_p}}{\sum_{i=1}^{T_p} e^{ll_i}}$$

   where p is the parameter (or parameter combination) used to produce the model output data, T_p is the total number of parameters drawn from the prior distribution and ll is the assigned log-likelihood value of the parameters.
4. The *acceptance criteria* for the BMLE method was thus that parameters had the chance of being accepted, equal (or given) the weights assigned to each parameter (or parameter combination). The parameters were then re-sampled with replacement using the `sample()` function in R (and the `sample_n()` function, in the `dplyr` package, for parameter combination re-sampling) according to the assigned weights, to produce a posterior distribution of parameters (see figure 2.4).

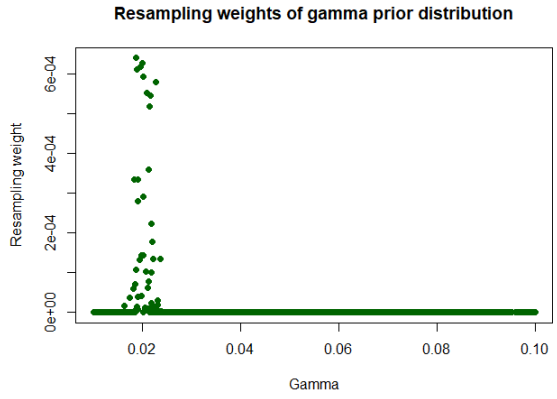
The result from each calibration using the BMLE method was thus a posterior distribution of the parameters from re-sampling parameters from the prior distribution according to the probability that the parameters were the target parameters.
The *stopping rule* for the BMLE method was thus when the sampling algorithm is completed. The time for the completion of the BMLE method thus depended on the number of parameters that were drawn from the prior distribution.



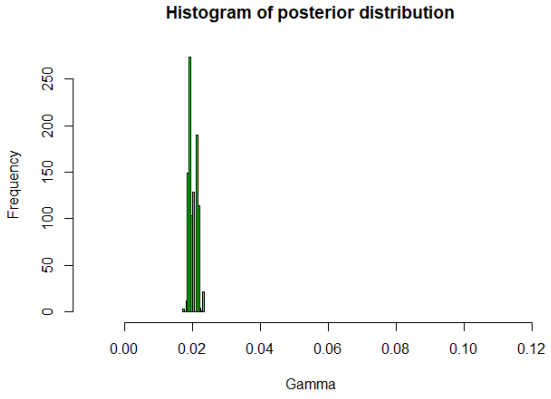
(a) Uniform prior distribution parameter plot



(b) Log-likelihood values of the prior distribution parameters



(c) The resampling weight values assigned to the prior distribution parameters



(d) Posterior distribution parameter plot

Figure 2.4: The Sampling importance resampling algorithm used in the BMLE method, where the uniform prior distribution is between $[0.01, 0.1]$ and the target parameter value of $\gamma = 0.02$

In every calibration using the BMLE method, 1000 parameters were randomly drawn from the prior distribution and 1000 parameters were re-sampled with replacement which formed the posterior distribution. The prior distributions from which parameters were randomly drawn had the form of a uniform distribution with bounds for $\beta = [0.01, 0.5]$ and $\gamma = [0.01, 0.1]$, similar to the starting value bounds for each parameter using the LS and MLE methods.

2.2.5.4 Approximate Bayesian Computation Rejection

The ABC-r method is also a Sampling method, which is the simplest form of the Approximate Bayesian Computation algorithms. The ABC-r method makes use of a rejection *parameter search strategy* (as seen in figure 2.5):

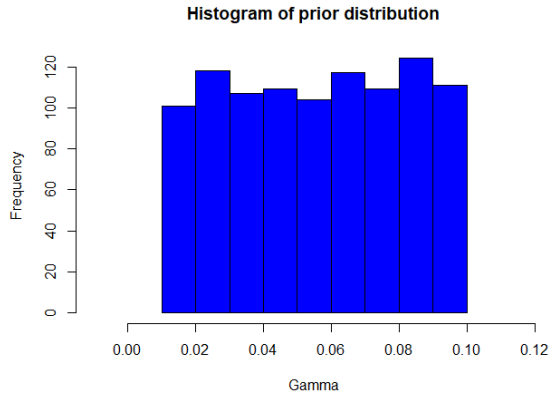
1. Parameters were randomly drawn from a uniform prior distribution, like the BMLE method.

2. Model output summary statistics were then produced for all the sampled parameters (or parameter combinations).

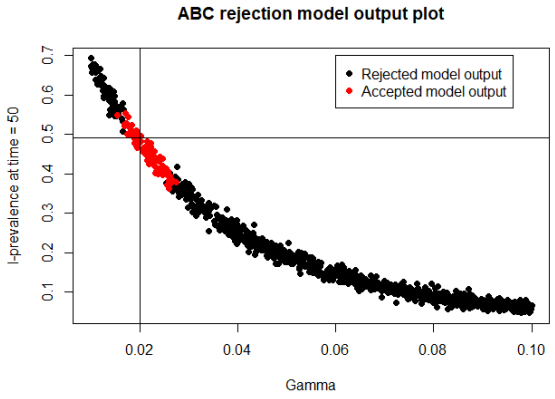
3. The Euclidean distance between the model output summary statistics using the parameters in the prior distribution and the target statistics were then calculated. This served as the *G.o.F* measure used.

4. From the prior distribution of parameters, the 10% best of the parameters that produced the smallest distances were then accepted and formed the posterior distribution of accepted parameters.

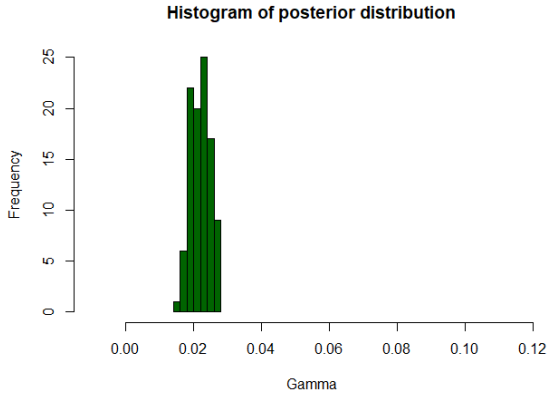
Thus the *acceptance criteria* of the ABC-r method was that parameters got accepted if they were part of a specified tolerance percentage of the parameters that produced the smallest Euclidean distance between the model output summary statistics and the target statistics (10% acceptance tolerance in this study, see 2.5b). The calibration result after using the ABC-r method was also a posterior distribution of parameters, as returned by the BMLE method.
The *stopping rule* for the ABC-r method was thus that when the ABC-r algorithm has completed, the calibration stops. As with the BMLE method, the time to complete the ABC-r method depended on the number of parameters drawn from the prior distribution.



(a) Uniform prior distribution parameter plot



(b) ABC rejection model output plot of infectious prevalence at time = 50, with 10% acceptance tolerance



(c) Accepted posterior distribution parameter plot

Figure 2.5: The ABC rejection algorithm, where the uniform prior distribution is between $[0.01, 0.1]$ and the target parameter value of $\gamma = 0.02$. In plot 2.5b the vertical line shows where the target parameter is and the horizontal line shows the target statistic at time point = 50

The ABC-r method made use of a uniform prior distribution and had randomly drawn 1000 parameters, like the BMLE method. The prior distributions were also specified the same as with the BMLE method. However, the ABC-r method had only retained the 10% best-fitting parameters, thus resulting in 100 parameters in the posterior distribution.

By specifying these bounds ensured that all the calibration methods had the same bounds for the explored parameters.

2.2.6 Performance Measures

The calibration methods were then compared to each other with the following performance measures: the average estimated parameter values over 1000 repeats, relative bias, accuracy (as the root mean square error) and coverage (the 95% confidence intervals).
These performance measures were necessary for testing the performance of the respective calibration in all the explored scenarios. This helped conclude which calibration methods performed better under which scenario and how the differences in the scenarios affected the performance of the calibration methods.

2.2.6.1 Average Estimated parameter

For every scenario, each of the calibration methods was executed 1000 times, i.e. 1000 pairs of target statistics were generated and for each of these 1000 pairs of targets, the calibration method was applied to estimate the target model parameters. This resulted in 1000 parameters (or parameter combinations) for each of the optimization methods and 1000 posterior distributions of parameters (or parameter combination) for each of the sampling methods, respectively.

Since the LS and MLE methods are optimization methods, they returned single value parameter estimates for every calibration model run performed. Thus by taking the mean of the 1000 parameter estimates respectively for each method, resulted in having an average parameter estimate for every explored scenario.

However, since BMLE and ABC-r are sampling methods, both methods returned posterior distributions of parameters for every calibration model run performed. For this reason, the median of the posterior distribution was used as a single value parameter estimate for each calibration model run, to be able to compare the results of the sampling methods to those of the optimization methods. The mean of the 1000 median parameter estimates were then taken for the two methods respectively, which thus resulted in also having an average parameter estimate for every explored scenario.

The average parameter estimates of the calibration methods were then used for further performance measuring, in every scenario.

2.2.6.2 Relative Bias

The calculation of relative bias helped to understand how far from the target parameter value, the calibration method had estimated the parameters. This was done by the following relative bias formula:

$$\text{Bias}_{\%} = \left(\frac{\overline{\hat{par}} - par)}{par} \right) \times 100$$

where par is the target parameter value and $\overline{\hat{par}}$ is the average estimated parameter value.
By using this formula, a percentage bias value was obtained, which also allowed for comparison between scenarios where more than one parameter had to be estimated since the β and γ parameters were not on the same scale.

The closer to 0% the percentage bias value, the less biased the calibration method was and the further from 0% the percentage bias value was, the more biased the calibration method was. When the percentage bias value was negative, the method underestimated the target parameter value and when the percentage bias was positive, the method overestimated the target parameter value.

2.2.6.3 Accuracy

The calculation of accuracy helped to understand how accurate the calibration method was able to estimate the target parameter values in each of the calibration model runs. By using the Root Mean Squared Error (RMSE) calculation, the evaluation of the accuracy of the calibration methods were achieved. The RMSE was thus calculated using the formula:

$$\text{RMSE} = \sqrt{\frac{\left[\sum_{i=1}^{n} (\hat{par}_i - par)^2\right]}{n}}$$

where n is the total number of estimated parameters returned from all the calibration model runs, par is the target parameter value and $\hat{par}$ is the estimated parameter value of each of the calibration model runs. In the accuracy calculation, the average of the estimated parameters was not used, thus the evaluation of the calibration methods accuracy indicates the performance of each of the 1000 calibration model runs, for the respective

calibration methods.

Since the RMSE was used, the mean deviation of all the estimated parameters from the target parameter was found. When the value was low, the calibration method had consistently estimated the parameters close to the target parameter value, thus the calibration method was more accurate. When the value was high, the calibration method consistently was unable to estimate parameters close to the target parameter value, thus being less accurate.

2.2.6.4 Coverage

The calculation of the coverage involved calculating the percentage of the 1000 estimated parameters that successfully had the target parameter value within their respective 95% confidence interval (CI). Thus to calculate the coverage for each of the calibration methods, the 95% CI had to be calculated for each of the 1000 estimated parameters from the respective calibrations.

Calculating the 95% CI involved first calculating the standard errors for each of the 1000 estimated parameters for the optimization methods. The `optim()` function, that the LS and MLE methods made use of has a built-in option to return a Hessian matrix for every estimated parameter result. From the respective Hessian matrices, the standard errors for all the respective estimated parameters were calculated, from which the 95% CI ranges were calculated by:

$$CI_{par}^{2.5} = \hat{par} - (1.96 \times SE)$$
$$CI_{par}^{97.5} = \hat{par} + (1.96 \times SE)$$

where $\hat{par}$ is the estimated parameter and SE is the standard error obtained from the Hessian matrix. Thus each estimated parameter had a 95% CI of the form $[CI_{par}^{2.5}, CI_{par}^{97.5}]$.

Since the median value of the posterior distribution from each calibration run was used as parameter estimates by the sampling methods, a 95% credible interval (CrI) was used to calculate the coverage of the BMLE and ABC-r methods. The posterior distributions were sorted from low to high values, then the value at the 2.5% index becomes the $CrI_{par}^{2.5}$ percentile value, and the value at the 97.5% index becomes the $CrI_{par}^{97.5}$ percentile value. Thus every median parameter estimate from the respective posterior distributions had a 95% CrI of the form $[CrI_{par}^{2.5}, CrI_{par}^{97.5}]$.

The coverage was then calculated using the following formula:

$$\text{Coverage} = \frac{\sum_{i=1}^{n} \left(\left(CI_{par_i}^{2.5} \leq par \leq CI_{par_i}^{97.5} \right) == TRUE \right)}{n} \times 100$$

where n was all the total number of estimated parameters that the calibration method produced and par was the target parameter value.

When the coverage was $< 95\%$, it meant that the 95% CI or CrI ranges of the estimated parameters were too narrow (the standard errors were small) and when the coverage was $> 95\%$ it meant that the 95% CI or CrI ranges of the estimated parameters were too wide (standard errors were big).

Chapter 3

Results

3.1 Introduction

This chapter contains the full scope of the results that were found in this study, by following the study design as described in Chapter 2.

Table 3.1 describes the differences between the four scenarios that were explored in this study. The scenarios were distinguished by looking at the combinations of two variables: sample population = 10% or 100% of the total population; and parameters to estimate = 1 (γ) or 2 (β, γ). In each of the scenarios, the four calibration methods were tested by having to estimate the scenario-specific amount of parameters using models that used 2, 3, 4 and 64 time points as target statistics of the observed data. The tables 3.2 to 3.5 contain the results for each of the scenarios, respectively.

Table 3.1: Scenarios for testing the calibration methods

Parameters	**Scenario 1**	**Scenario 2**	**Scenario 3**	**Scenario 4**
Sample size (% of Total Population)	10%	100%	10%	100%
Parameters to estimate	γ	γ	(β, γ)	(β, γ)

3.2 Main Findings

3.2.1 Performance of calibration methods within scenarios

Here the results of the four calibration methods are compared to each other within the four scenarios.

Table 3.2: Scenario 1: one parameter, sample size = 10%

Target Statistics	Calibration Method	$\bar{\hat{\gamma}}$	Percentage bias	RMSE	Coverage
2 Targets at time = 50, 65	LS	0.034	70%	0.028	84.7%
	MLE	0.032	60%	0.026	56.7%
	BMLE	0.02	0%	0	100%
	ABC-r	0.021	5%	0.001	100%
3 Targets at time = 50, 65, peak prev	LS	0.031	55%	0.024	84.9%
	MLE	0.031	55%	0.025	60.6%
	BMLE	0.02	0%	0	100%
	ABC-r	0.021	5%	0.001	100%
4 Targets at time = 30, 45, 60, 75	LS	0.031	55%	0.025	84.1%
	MLE	0.032	60%	0.026	59.6%
	BMLE	0.02	0%	0	100%
	ABC-r	0.021	5%	0.001	100%
64 Targets at time = 1:64	LS	0.024	20%	0.015	94%
	MLE	0.024	20%	0.015	45.1%
	BMLE	0.02	0%	0.001	30.2%
	ABC-r	0.02	0%	0.001	100%

Table 3.3: Scenario 2: one parameter, sample size = 100%

Target Statistics	Calibration Method	$\bar{\hat{\gamma}}$	Percentage bias	RMSE	Coverage
2 Targets at time = 50, 65	LS	0.028	40%	0.023	92%
	MLE	0.028	40%	0.022	64.9%
	BMLE	0.02	0%	0	99.2%
	ABC-r	0.021	5%	0.001	100%
3 Targets at time = 50, 65, peak prev	LS	0.023	15%	0.015	96.6%
	MLE	0.024	20%	0.016	65.7%
	BMLE	0.02	0%	0	97.6%
	ABC-r	0.021	5%	0.001	100%
4 Targets at time = 30, 45, 60, 75	LS	0.025	25%	0.017	93%
	MLE	0.025	25%	0.018	32.5%
	BMLE	0.02	0%	0	95.6%
	ABC-r	0.021	5%	0.001	100%
64 Targets at time = 1:64	LS	0.023	15%	0.013	95.7%
	MLE	0.023	15%	0.015	21%
	BMLE	0.02	0%	0	14.2%
	ABC-r	0.02	0%	0.001	100%

3.2.1.1 Scenario 1: Estimating one parameter (γ) with sample size = 10% of total population

In scenario 1 (as seen in table 3.2) the calibration methods were tasked to only estimate the target parameter $\gamma = 0.02$, with the sample size being 10% of the total population, $N = 10000$.

The LS and MLE methods had similar performance measure values, with the higher

Table 3.4: Scenario 3: two parameters, sample size = 10%

Target Statistics	Calibration Method	$(\overline{\hat{\beta}}, \overline{\hat{\gamma}})$	Percentage bias	RMSE	Coverage
2 Targets at time = 50, 65	LS	(0.289, 0.029)	(44.5%, 45%)	(0.187, 0.025)	(31.5%, 86.2%)
	MLE	(0.289, 0.029)	(44.5%, 45%)	(0.191, 0.025)	(0.4%, 15.6%)
	BMLE	(0.281, 0.019)	(40.5%, −5%)	(0.096, 0.002)	(99.7%, 99.6%)
	ABC-r	(0.253, 0.02)	(26.5%, 0%)	(0.058, 0.001)	(100%, 100%)
3 Targets at time = 50, 65, peak prev	LS	(0.319, 0.03)	(59.5%, 50%)	(0.193, 0.025)	(23.6%, 83.8%)
	MLE	(0.252, 0.057)	(26%, 185%)	(0.19, 0.049)	(0.8%, 3.6%)
	BMLE	(0.205, 0.02)	(2.5%, 0%)	(0.02, 0.001)	(75.9%, 77.3%)
	ABC-r	(0.263, 0.02)	(31.5%, 0%)	(0.066, 0.001)	(100%, 100%)
4 Targets at time = 30, 45, 60, 75	LS	(0.306, 0.027)	(53%, 35%)	(0.187, 0.022)	(31.8%, 86.8%)
	MLE	(0.304, 0.025)	(52%, 25%)	(0.194, 0.025)	(1%, 14.9%)
	BMLE	(0.316, 0.018)	(58%, −10%)	(0.135, 0.002)	(64.5%, 70.8%)
	ABC-r	(0.276, 0.02)	(38%, 0%)	(0.079, 0.001)	(100%, 100%)
64 Targets at time = 1:64	LS	(0.308, 0.024)	(54%, 20%)	(0.172, 0.016)	(23.9%, 89.6%)
	MLE	(0.306, 0.023)	(53%, 15%)	(0.177, 0.014)	(1.7%, 19.2%)
	BMLE	(0.201, 0.02)	(0.5%, 0%)	(0.01, 0.001)	(4.7%, 5.6%)
	ABC-r	(0.252, 0.02)	(26%, 0%)	(0.055, 0.001)	(100%, 100%)

Table 3.5: Scenario 4: two parameters, sample size = 100%

Target Statistics	Calibration Method	$(\overline{\hat{\beta}}, \overline{\hat{\gamma}})$	Percentage bias	RMSE	Coverage
2 Targets at time = 50, 65	LS	(0.295, 0.023)	(47.5%, 15%)	(0.185, 0.017)	(76.4%, 95.7%)
	MLE	(0.255, 0.055)	(27.5%, 175%)	(0.188, 0.048)	(0.7%, 4.6%)
	BMLE	(0.239, 0.019)	(19.5%, −5%)	(0.075, 0.002)	(84.7%, 84.6%)
	ABC-r	(0.253, 0.02)	(26.5%, 0%)	(0.057, 0.001)	(100%, 100%)
3 Targets at time = 50, 65, peak prev	LS	(0.316, 0.022)	(58%, 10%)	(0.183, 0.014)	(45%, 97.1%)
	MLE	(0.301, 0.022)	(50.5%, 10%)	(0.183, 0.014)	(2.1%, 28.8%)
	BMLE	(0.202, 0.02)	(1%, 0%)	(0.013, 0.001)	(15.4%, 15.5%)
	ABC-r	(0.263, 0.02)	(31.5%, 0%)	(0.066, 0.001)	(100%, 100%)
4 Targets at time = 30, 45, 60, 75	LS	(0.315, 0.021)	(57.5%, 5%)	(0.185, 0.014)	(62.8%, 95.9%)
	MLE	(0.316, 0.022)	(58%, 10%)	(0.195, 0.015)	(1.5%, 7.6%)
	BMLE	(0.266, 0.019)	(33%, −5%)	(0.102, 0.002)	(28.2%, 29.1%)
	ABC-r	(0.276, 0.02)	(38%, 0%)	(0.078, 0.001)	(100%, 100%)
64 Targets at time = 1:64	LS	(0.299, 0.023)	(49.5%, 15%)	(0.167, 0.013)	(27.4%, 92.4%)
	MLE	(0.309, 0.023)	(54.5%, 15%)	(0.176, 0.013)	(0.5%, 7.6%)
	BMLE	(NAN, NAN)	(NAN, NAN)	(NAN, NAN)	(NAN, NAN)
	ABC-r	(0.252, 0.02)	(26%, 0%)	(0.054, 0.001)	(100%, 100%)

values of percentage bias and RMSE values than BMLE and ABC-r in this scenario. The MLE method had lower coverage values than the LS method throughout this scenario. The BMLE method had an exact mean parameter estimate of $\overline{\hat{\gamma}} = 0.02$ for all four number of target statistics, while also having RMSE values of 0 at 2, 3 and 4 target statistics and 0.001 at 64 target statistics. At 64 target statistics, the BMLE method had a deviating coverage value of 30.2%, whereas it had 100% at the other number of target statistics. The ABC-r method had the same performance measure values at all the number of target statistics, except it had a percentage bias value of 0% at 64 target statistics, other than the 5% found at the other number of target statistics.

3.2.1.2 Scenario 2: Estimating one parameter (γ) with sample size = 100% of total population

In scenario 2 (as seen in table 3.3) the calibration methods were tasked to again only estimate the target parameter $\gamma = 0.02$, however with the sample size being 100% of the total population, $N = 10000$.

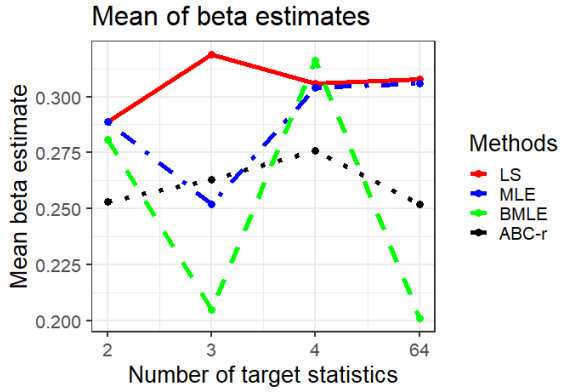
The LS and MLE methods again had similar performance measure values in this scenario, also with the MLE method having lower coverage values than the LS method, as seen at 64 target statistics where the LS method had a coverage value of 95.7% and the MLE method had 21%.
The BMLE method had percentage bias and RMSE values of 0 at all the number of target statistics. At 64 target statistics, the BMLE method, however, had a coverage value of 14.2%, a lower value than at the other number of target statistics.
The ABC-r method again had the same performance measure values at all the number of target statistics, while having a percentage bias values of 0% at 64 target statistics, other than the 5% found at the other number of target statistics,

3.2.1.3 Scenario 3: Estimating two parameters (β, γ) with sample size = 10% of total population

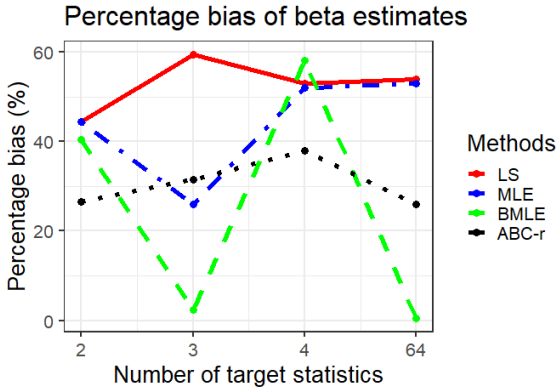
In scenario 3 (as seen in table 3.4) the calibration methods were tasked to estimate two target parameters: $\beta = 0.2$ and $\gamma = 0.02$. With the sample size being 10% of the total population, $N = 10000$. The results of this scenario are also highlighted in figures 3.1 and 3.2, since the variables in this scenario is more likely to be that of a real-world study.

At 3 target statistics, the MLE method had a percentage bias value of 26% for β but 185% for γ, with a RMSE value of 0.049 for γ. The performance measure values of β and γ were similar between the LS and MLE methods at 2, 4 and 64 target statistics, except for the coverage values, where the MLE method had lower values. The highest coverage value of β that the MLE method had was 1.7% at 64 target statistics, whereas the LS method had a coverage of 23.9%, its second to lowest.
The BMLE method had a mean parameter estimate of $\overline{\hat{\beta}} = 0.201$ at 64 target statistics, but had a coverage value of 4.7%. At 4 target statistics, the BMLE method had a percentage bias value for γ of -10%.
The ABC-r method had the same performance measure values for γ at all of the number of target statistics, with percentage bias values of 0% and RMSE values of 0.001. The

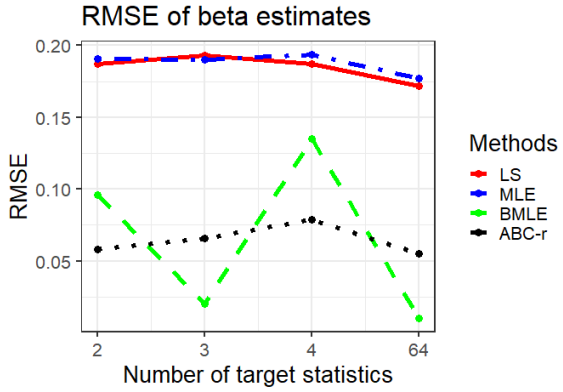
ABC-r method had its lowest percentage bias value for β of 26% at 64 target statistics.



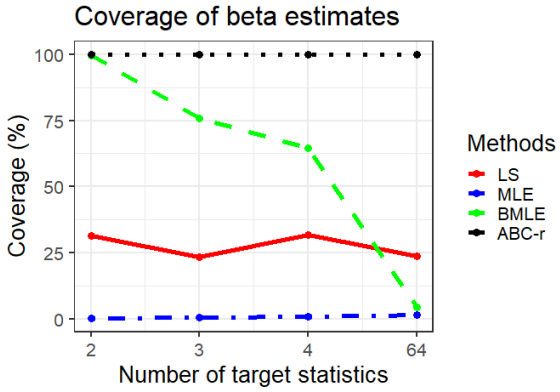
(a) $\bar{\hat{\beta}}$ of all the calibration methods at the different numbers of target statistics



(b) The percentage bias of β of all the calibration methods at the different numbers of target statistics



(c) The RMSE value of β of all the calibration methods at the different numbers of target statistics

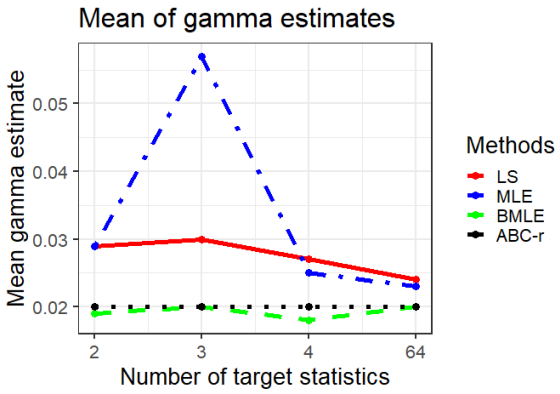


(d) The coverage value of β of all the calibration methods at the different numbers of target statistics

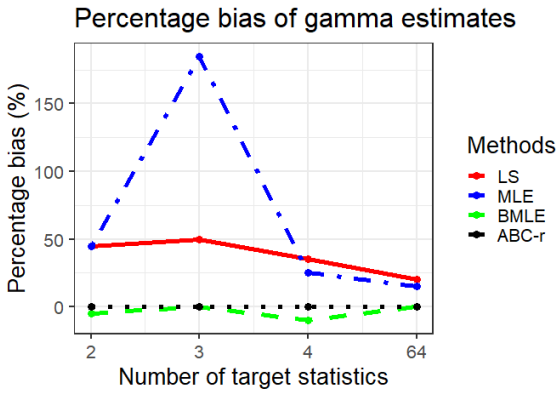
Figure 3.1: The performance measures values of the estimation of the target parameter β using the calibration methods in scenario 3

3.2.1.4 Scenario 4: Estimating two parameters (β, γ) with sample size = 100% of total population

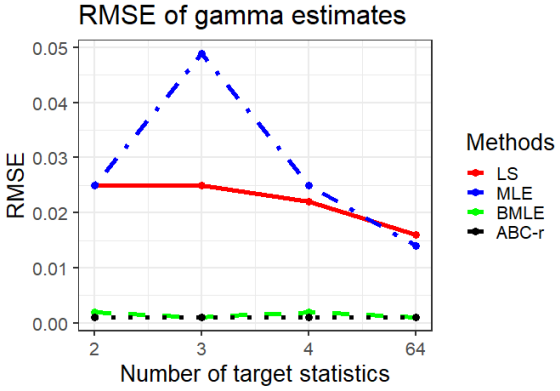
In scenario 4 (as seen in table 3.5) the calibration methods were again tasked to estimate two target parameters: $\beta = 0.2$ and $\gamma = 0.02$. However, with the sample size being 100% of the total population, $N = 10000$.



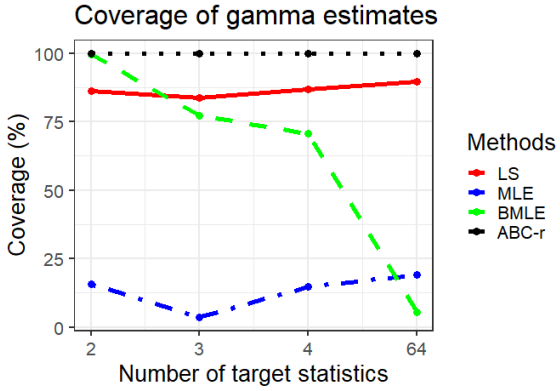
(a) $\overline{\hat{\gamma}}$ of all the calibration methods at the different numbers of target statistics



(b) The percentage bias of γ of all the calibration methods at the different numbers of target statistics



(c) The RMSE value of γ of all the calibration methods at the different numbers of target statistics



(d) The coverage value of γ of all the calibration methods at the different numbers of target statistics

Figure 3.2: The performance measures values of the estimation of the target parameter γ using the calibration methods in scenario 3

The MLE method had a percentage bias value for γ of 175% at 2 target statistics, with a RMSE value of 0.048. The MLE method also had a percentage bias value for γ of 27.5% at 2 target statistics. At all the other number of target statistics the LS and MLE methods again had similar performance measure values, except for the coverage, which was lower for the MLE method.

At 3 target statistics the BMLE method had a percentage bias value for β and γ of 1% and 0% respectively, but with coverage values of 28.2% for β and 29.1% for γ. At 64 target statistics the BMLE method produced unusable weight values, resulting in NAN vales of parameter estimates for β and γ.

The ABC-r method again had the same performance bias values for γ at all the number of target statistics, with percentage bias values of 0% and RMSE values of 0.01.

Chapter 4

Discussion

4.1 Discussion

4.1.1 Performance of the calibration methods

To conclude which calibration method performed the best in each scenario comes down to defining how the performance measures define the best performance. The performance results can thus be interpreted by three criteria; for most of the different number of target statistics, either:

1. the calibration method that had the least bias values performed the best,
2. the calibration method that had the best combination of accuracy and coverage performed the best or
3. the calibration method that had the best combination of all the performance measures performed the best.

These different performance criteria comes down to the interpretation that either reducing bias or reducing variability increases precision in parameter estimation.
Thus to find the calibration method that performed the best in each scenario, table 4.1 was constructed which counts the number of times (for how many number of target statistics) the best performance measure values of the calibration methods were found, per criteria. For percentage bias and RMSE, the method that produced the lowest value received 1 score for that target statistic and coverage, the method that produced the smallest absolute deviation from 95% (closest to 95%) received 1 score for that target statistic. However, when more than one method found the same best value, no score was given to either of the methods for that target statistic.

Table 4.1: The scores of the number of times the calibration methods had the best performance measure value at every number of target statistics per scenario. When methods had the same performance measure value, a 0 value was given for that performance measure at the specific target statistic.

Scenario (parameters)	Calibration method	Percentage bias	RMSE	Coverage
1 (γ)	LS	0	0	1
	MLE	0	0	0
	BMLE	3	3	0
	ABC-r	0	0	0
2 (γ)	LS	0	0	3
	MLE	0	0	0
	BMLE	3	4	1
	ABC-r	0	0	0
3 (β, γ)	LS	(0, 0)	(0, 0)	(0, 0)
	MLE	(0, 0)	(0, 0)	(0, 0)
	BMLE	(2, 0)	(2, 0)	(1, 1)
	ABC-r	(2, 4)	(2, 2)	(3, 3)
4 (β, γ)	LS	(0, 0)	(0, 0)	(0, 4)
	MLE	(0, 0)	(0, 0)	(0, 0)
	BMLE	(3, 0)	(1, 0)	(0, 0)
	ABC-r	(1, 2)	(3, 3)	(4, 0)

Thus, given the information from table 4.1, the best calibration method for the different criteria per scenarios are found. For criteria 2 and 3 the scores of the specified performance measures are combined and for scenarios 3 and 4 the scores for β and γ are also combined, for every calibration method. The best calibration methods, given the scores and different criteria, per scenarios was thus:

1. Scenario 1
    - Criteria 1: BMLE
    - Criteria 2: BMLE
    - Criteria 3: BMLE
2. Scenario 2
    - Criteria 1: BMLE
    - Criteria 2: BMLE
    - Criteria 3: BMLE

3. Scenario 3
    - Criteria 1: ABC-r
    - Criteria 2: ABC-r
    - Criteria 3: ABC-r
4. Scenario 4
    - Criteria 1: BMLE and ABC-r
    - Criteria 2: ABC-r
    - Criteria 3: ABC-r

Thus, even though different criteria were specified for the interpretation of the best calibration method, it can be concluded that the BMLE method had done the best in scenarios 1 and 2 and the ABC-r method had done the best in scenario 3 and 4 (even though for criteria 1, the scores for the BMLE and ABC-r methods were the same, the ABC-r method had the overall best performance).

4.1.2 Effects on the performance of calibration methods when changing key variables

The following section discusses the impact the changes in key variables had. By intra-scenario inspection of the performance measure results for each calibration method, the impact of increasing the number of target statistics from 2 to 64 can be evaluated. By inter-scenario inspection between scenarios 1 and 2 and between scenarios 3 and 4 the impact of increasing the sample size from 10% to 100% of the total population can be evaluated since it was the only difference between these scenarios. Also, by inter-scenario inspection between scenarios 1 and 3 and between scenarios 2 and 4 of the performance measure values of γ, the impact of increasing the number of parameters to estimate from 1 to 2 can be evaluated, since it was the only difference between these scenarios.

4.1.2.1 Least-Squares

The increase in the number of target statistics mostly had an impact bias and RMSE values. Not much of an impact was observed for the coverage values. Bias and RMSE values of γ decreased as the number of target statistics increased, however in scenarios 3 and 4 the bias and RMSE values of β did not decrease much.

The increase in the sample size had an impact on all of the performance measure values. There were clear improvements for the γ performance measure values but β the bias and RMSE values did not improve as much. As seen between scenarios 3 and 4, where the biggest improvement was at 64 target statistics where the bias value for β was 54% and RMSE value was 0.172 in scenario 3 and in scenario 4 the bias value was 49.5 and the RMSE values was 0.167. At 4 target statistics, however, there was an increase in bias and RMSE values for β between these two scenarios.

The increase in the number of parameters to estimate mostly had an impact on the bias values of γ at 2, 3 and 4 target statistics. At 64 target statistics, the performance measure values were very similar for γ between the scenarios, and the coverage value in scenario 1 was better than the coverage value in scenario 3.

4.1.2.2 Maximum Likelihood Estimation

The increase in the number of target statistics most had an impact on the bias and RMSE values of γ, where these values decreased as the number of target statistics increased. However, the coverage values for both γ and β did not improve with the increase in target statistics, where the coverage values for γ decreased as the number of target statistics increased.

Increasing the sample size improved the bias and RMSE values for γ, but the coverage values were similar between scenarios 1 and 2 and between scenarios 3 and 4 varying effects were observed. Therefore the increase in sample size did not have much of an improvement in coverage values. For β the increase in the sample size did not have much of an improvement in performance measure values.

The increase in parameters to estimate only improved the bias values for γ, the RMSE values remained similar and the coverage values drastically decreased. It has to be noted that with the increase in parameters to estimate between scenario 1 and 3, at 3 target statistics the bias values drastically increased from 55% to 185% and between scenario 2 and 4, at 2 target statistics the bias values increased from 40% to 175%. This could be an indication of the limitations of the Nelder-mead method (used in the `optim()` function), where the method usually finds local minima (depending on the starting values of the search) and not the absolute minima. Further investigation of this issue could be done in a further study.

4.1.2.3 Bayesian Maximum Likelihood Estimation

The increase in the number of target statistics did not have an impact on the bias and RMSE for γ but it did have a decreasing effect on the coverage values of γ. For β the increase in the number of target statistics had decreased bias, RMSE and coverage values but resulted in NAN values for all the performance measure values at 64 target statistics in scenario 4. With the scope of the scenario at 64 target statistics, the likelihood function of the BMLE method produced very small log-likelihood values, which resulted in INF and 0 values in the weight calculation step of the method, thus resulting in no values produced during the measuring of the performance of the method.

Increasing the sample size also had no impact on the bias and RMSE values of γ but it had decreased the coverage values. The increase in sample size had improved the performance measure values for β except for the coverage values, which decreased.

Increasing the number of parameters to estimate had a minor impact on the bias values for γ where between scenarios 1 and 3, at both 2 and 4 target statistics the bias values went from 0% to -5% and -10% respectively. A minor increase in RMSE values and a decrease in coverage values was also observed as the number of parameters to estimate increased.

4.1.2.4 Approximate Bayesian Computation - rejection

The increase in the number of target statistics had no impact on the performance measure values of γ. A slight increase in bias and RMSE values of β was observed between 2, 3 and 4 target statistics but at 64 target statistics, these values were slightly less than the values at 2 target statistics. The coverage values consistently remained 100% for both β and γ. This could be an indication of very wide CrI ranges around the individual parameter estimates.

The increase in sample size also had no impact on the performance measure values of β and γ.

Increasing the number of parameters to estimate also had no impact on the performance measure values of γ.

Chapter 5

Conclusion

5.1 Conclusion

This study aimed to evaluate the performance and compare four calibration methods to each other under different scenarios. This was achieved by implementing a simulation study where a simple stochastic SIR model was calibrated to simulated data using the calibration methods and evaluating their performance using three different performance measures.

It was found that sampling methods performed the best calibrations by producing parameter estimates that minimized bias, maximized accuracy and found sufficient coverage of the target parameters of the simulated data. More specifically it was found that the BMLE calibration method performed the best when the only parameter had to be estimated and the ABC-r method performed the best when two parameters had to be estimated.

It was also found that the change in number of target statistics, sample size and number of parameters to estimate only had an impact on the optimization methods. By increasing the number of target statistics and sample size the LS and MLE methods produced lower percentage bias values, however, the increase in the number of parameters to estimate from 1 to 2 did not have as big of an impact on the performance of the LS and MLE methods.
The sampling methods produced similar performance measure results even when changing these variables, however in using the BMLE method with the maximum values of these variables (64 target statistics, 10000 sample size and 2 parameters to estimate) re-

sulted in NAN values for all the performance measure values of the method. The conclusion on the coverage of BMLE, it decreases with increasing number of target statistics. Which is an interesting result, this most likely occurs because we do not account for the correlation between consecutive targets in our likelihood specification.

5.2 Limitations and Strengths

This study has found very interesting results and conclusions but given the study design, a few limitations do exist.
The study made use of a simple SIR model from which the calibration only attempted to estimate a maximum of 2 parameters. The study could thus have found different results if a more complicated model was used and the number of parameters to estimate was increased. As seen in the results, when increasing the number of parameters to estimate from 1 to 2, the LS and MLE methods had slight improvements in results, however, it is not clear yet how much of this might impact the precision of these calibration methods. Also because of the chosen mathematical model, there were not many options for different types of summary statistics. The infectious curve time points have provided meaningful outcome responses to different values of the estimated parameter, but a more complicated model might have produced more meaningful response variables to the changes in different parameters.
The number of model runs may also have affected the results of this study. Since only 1000 were used per calibration, it leaves some questions to what impact increasing this number may have given, however, given the computation time required to have run these calibrations, 1000 model runs were sufficient for this study. Also, the total population and sample sizes used in this study could have been increased but again the values used in this study were because of limitations in computation time and power. The sampling methods, BMLE and ABC-r had very long running times per calibration and are very computationally intensive methods, which made it difficult to explore the impact of increasing these values.

Despite these limitations, this study provided a lot of insight into the strengths of different calibration methods as well as where they fall short. This study provided a good framework in which calibration methods can be compared by being implemented as a simulation study. By the specification of the same bounds for the explored parameters and by evaluating the calibration methods using well-known performance measures, it also provided a good framework for the comparison of methods. This study has pro-

vided the insights into which calibrations methods give parameter estimates with the most uncertainties (most deviations from the target parameters) as well as the least uncertainties (minimize bias and maximize accuracy and coverage).

5.3 Future Research

This study provides a good basis for future research on calibration methods to be conducted. This study can be improved on in a few and simple ways by:

- using different models and comparing how different calibration methods perform on these models,
- using different types of summary statistics, which can also explain the impact of the choice of target model output data,
- increasing the number of parameters to explore,
- using higher values of model runs and random prior distribution draws, etc.

Appendix A

Appendix

A.1 Code

The following is a link to access a GitHub repository to find the code used to perform this study:
Wynand-R-Code.

List of references

Aldrich, J. (1997). R. a. fisher and the making of maximum likelihood 1912 - 1922. *Statistical Science*, vol. 12, no. 3, pp. 162 – 176.

Andrews, I. and Mijusheva, A. (). Weak identification in maximum likelihood: A question of information.

Beaumont, M.A. (2010 August). Approximate bayesian computation in evolution and ecology. *Annual Review of Ecology, Evolution and Systematics*, vol. 41, pp. 379–406.

Beaumont, M.A., Zhang, W. and Balding, J.D. (2002 December). Approxmate bayesian computation in popluation genetics. *Genetics Society of America*, vol. 162, no. 4, pp. 2025–2035.

Beerli, P. (2006). Comparison of bayesian and maximum-likelihood inference of population genetic parameters. *Bioinformatics*, vol. 22, no. 3, pp. 341–345.

Briggs, A.H., Weinstein, M.C., Fenwick, E.A.L., Karnon, J., Sculpher, M.J. and Paltiel, A.D. (2012). Model parameter estimation and uncertainty analysis: A report of the ispor-smdm modeling good research practices task force working group -6. *Mdeical Decision Making*.

Burton, A., Altman, D.G., Royston, P. and Holder, R.L. (2006 August). The design of simulation studies in medical statistics. *Statistics in Medicine*, vol. 25, pp. 4279–4292.

Byrd, R.H., Lu, P., Nocedal, J. and Zhu, C. (1994). A limited memory algorithm for bound constrained optimization. Tech. Rep., Northwestern Univerity.

Chowell, G. (2017). Fitting dynamic models to epidemic outbreaks with quantified uncertainty: A primer for parameter uncertainty, identifiability, and forecasts. *Infectious Disease Modelling*, vol. 2, no. 3, pp. 379 – 398. ISSN 2468-0427.
Available at: http://www.sciencedirect.com/science/article/pii/S2468042717300234

Collins, M. (). The naive bayes model, maximum-likelihood estimation, and the em algorithm.

Dahabreh, I.J., Chan, J.A., Earley, A., Moorthy, D., Avendano, E.E., Trikalinos, T.A., Balk, E.M. and Wong, J.B. (2017 January). *Modeling and Simulation in the Context of Health Technology Assessment: Review of Existing Guidance, Future Research Needs, and Validity Assessment*. Agency for Healthcare Research and Quality.

Dalzell, C. (2013 October). Optimization in r. Webpage. Viewed on 29/10/2019. Available at: https://www.ibm.com/developerworks/library/ba-optimR-john-nash/index.html

Fienberg, S.E. (2006). When did bayesian inferencen become "bayesian"? *International Society for Bayesian Analysis*, vol. 1, no. 1, pp. 1–40.

Gavin, H.P. (2016). The nelder-mead algorithm in two dimensions. Tech. Rep., Duke Univeristy.

Gillespie, D.T. (1977 May). Exact stochastic simulation of coupled chemical reactions. *Jounrnal of Physical Physics*, vol. 81, no. 25.

Jackson, C.H., Jit, M., Sharples, L.D. and Daniela De Angelis, P. (2015). Calibration of complex models through bayesian evidence synthesis: A demonstration and tutorial. *Medical Decision Making*.

Johnston, A.S.A., Hodson, M.E., Thorbek, P., Alvarez, T. and Sibly, R.M. (2014). An energy budget agent-based model of earthworm populations and its application to study the effects of pesticide. *Ecological Modelling*, vol. 280, pp. 5–17.

Karnon, J. and Vanni, T. (2011). Calibration model in economic evaluation: A comparison of alternative measures of goodness of fit, parameters search strategies and convergence criteria. *Pharmacogenetics*, vol. 29, no. 1, pp. 51–62.

Kong, C.Y., McMahon, P.M. and Gazelle, G.S. (2009). Calibrating of disease simulation model using an engineering approach. *Value in Health*, vol. 12, no. 4, pp. 521–529.

Menzies, N.A., Soeteman, D.I., Pandya, A. and Kim, J.J. (2017 Jun). Bayesian methods for calibrating health policy models: A tutorial. *PharmacoEconomics*, vol. 35, no. 6, pp. 613–624. ISSN 1179-2027.
Available at: https://doi.org/10.1007/s40273-017-0494-4

Myung, J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, vol. 47, pp. 90–100.

Nelder, J.A. and Mead, R. (1965 01). A simplex method for function minimization. *The Computer Journal*, vol. 7, no. 4, pp. 308–313. ISSN 0010-4620. http://oup.prod.sis.lan/comjnl/article-pdf/7/4/308/1013182/7-4-308.pdf.
Available at: https://doi.org/10.1093/comjnl/7.4.308

Pernot, P. and Calliez, F. (2017 April). A critcial review of statistical calibration/predection models handling data inconsistnecy and model inadequacy.

Pineda-Krch, M. (2008). Gillespiessa: Implementing the stochastic simulation algorithm in r. *Journal of Statistical Software*, vol. 25, no. 12.

Pitt, M.A. and Myung, I.J. (2002 October). When a good fit can be bad. *Trends in Cognitive Sciences*, vol. 6, no. 10, pp. 421–425.

Punyacharoensin, N., Edmunds, W.J., De Angelis, D. and White, R.G. (2011 September). Mathematical models for the study of hiv spread and control amongst men who have sex with men. *European Jounral of Epidemiolgy*, vol. 26, pp. 695–709.

Sisson, S.A., Fan, Y. and Beaumont, M.A. (2018 February). Overview of approximate bayesian computation.

Sorenson, H.W. (1970 July). Least-squares estimation: from gauss to kalman. *IEEE spectrum*.

Stout, N., Knusden, A., Kong, C., McMahon, P. and Gazelle, G. (2009). Calibration methods used in cancer simulation models and suggested reporting guidelines. *Pharmacoeconomics*, vol. 27, no. 7, pp. 533–545.

Taylor, D.C., Pawar, V., Kruzikas, D., Gilmore, K.E., Pandya, A., Iskandar, R. and Weinstein, M.C. (2010). Methods of model calibration: Observation from a mathematical model of cervical cancer. *Pharmacoeconomics*, vol. 28, no. 11, pp. 995–1000.

Taylor, D.C.A., Pawar, V., Kruzikas, T., Gilmore, K.E., Sanon, M. and Weinstein, M.C. (2012). Incorporating calibrated model parameters into sensitivty analyses: Detemi-nistic and probabilistic approaches. *Pharmacoeconomics*, vol. 30, no. 2, pp. 119–126.

Van De Geer, S.A. (2005). Least squares estimation. *Encyclopedia of Statistics in Behavioral Science*, vol. 2, pp. 1041–1045.

van der Vaart, E., Beaumont, M.A., Johnstona, A.S. and Sibly, R.M. (2015). Calibration and evaluation of individual-based models using approximate bayesian computation. *Ecoloical Modelling*, vol. 312, pp. 182–190.

van Smeden, M., de Groot, J.A.H., Moons, K.G.M., Collins, G.S., Altman, D.G., Eijkemans, M.J.C. and Reitsma, J.B. (2016). No rationale for 1 vairable per 10 events criterion for binary logistic regression analysis. *BMC Medical Research Methodology*, vol. 16, no. 163.

Vanni, T., Karnon, J., Madan, J., White, R.G., Edmunds, W.J., Foss, A.M. and Legood, R. (2011). Calibrating models in economic evaluation: A seven-step approach. *Pharmacoeconomics*, vol. 29, no. 1, pp. 35–49.

Widgren, S., Bauer, P. and Engblom, S. (2016). Siminf: An r package for data-driven stochastic disease spread simulations. Tech. Rep., Uppsala University, Sweden.

Widgren, S., Eriksson, R., Engblom, S. and Bauer, P. (2019 November). Package 'siminf': A framework for data-driven stochastic disease spread simulations. Tech. Rep., Uppsala University, Sweden. Version 6.4.0.

Wilkinson, R. (2008). Approximate bayesian computation (abc) gives exact results under the assumption of model error. *Biometrika*.